L'accès au lexique dans la perception audiovisuelle et visuelle de la parole

Mathilde Fort

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Sciences cognitives, Psychologie et Neurocognition**

Arrêté ministériel : 7 août 2006

Présentée par

Mathilde FORT

Thèse dirigée par **Sonia KANDEL** et co-dirigée par **Elsa SPINELLI**

Préparée au sein du **Laboratoire de Psychologie et NeuroCognition – CNRS UMR 5105**

Dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

L'accès au lexique dans la perception audiovisuelle et visuelle de la parole

Thèse soutenue publiquement le **5 Décembre 2011**,
devant le jury composé de :

Pascal, BARONE
Directeur de Recherche CNRS, CERCO, Université Paul Sabatier de Toulouse, Rapporteur

Juan, SEGUI
Directeur de Recherche Emérite CNRS, LPNCog, Université Paris Descartes, Paris, Rapporteur

Pierre, HALLE
Directeur de Recherche à l'Université Sorbonne Nouvelle, Paris, Examinateur et Président du jury

Ulrich, FRAUENFELDER
Professeur à l'Université de Genève, Examinateur

Jean-Luc, SCHWARTZ
Directeur de Recherche CNRS, Gipsa-lab, Grenoble, Examinateur

Sonia, KANDEL
Professeur à l'Université de Grenoble, LPNC, CNRS, Grenoble, Directrice de thèse

Elsa, SPINELLI
Maître de Conférences HDR à l'Université de Grenoble, LPNC, CNRS UMR 5105, co-directrice de thèse

A Mamy de Juvigny

A Papy de Saillans

Résumé

En situation de perception audiovisuelle de la parole (i.e., lorsque deux interlocuteurs communiquent face à face) et lorsque le signal acoustique est bruité, l'intelligibilité des sons produits par un locuteur est augmentée lorsque son visage en mouvement est visible. L'objectif des travaux présentés ici est de déterminer si cette capacité à « lire sur les lèvres » nous est utile seulement pour augmenter l'intelligibilité de certains sons de parole (i.e., niveau de traitement pré-lexical) ou également pour accéder au sens des mots (i.e., niveau de traitement lexical). Chez l'adulte, nos résultats indiquent que l'information visuelle participe à l'activation des représentations lexicales en présence d'une information auditive bruitée (Etude 1 et 2). Voir le geste articulatoire correspondant à la première syllabe d'un mot constitue une information suffisante pour contacter les représentations lexicales, en l'absence de toute information auditive (Etude 3 et 4). Les résultats obtenus chez l'enfant suggèrent néanmoins que jusque l'âge de 10 ans, l'information visuelle serait uniquement décodée à un niveau pré-lexical (Etude 5).

Mots-clés : parole visuelle et audiovisuelle, reconnaissance de mots parlés, accès au lexique.

Abstract

Seeing the facial gestures of a speaker enhances phonemic identification in noise. The goal of this research was to assess whether this visual information can activate lexical representations. We investigated this question in adults (Experiment 1 to 4) and in children (Experiment 5). First, our results provide evidence indicating that visual information on consonant (Experiment 1) and vowel identity (Experiment 2) contributes to lexical activation processes during word recognition, when the auditory information is deteriorated by noise. Then, we also demonstrated that the mere presentation of the first two phonemes – i.e., the articulatory gestures of the initial syllable– is enough visual information to activate lexical representations and initiate the word recognition process (Experiment 3 and 4). However, our data suggest that visual speech mostly contributes in pre-lexical phonological -rather than lexical- processing in children till the age of 10 (Experiment 5).

Key words : speech, visual and audiovisual speech, spoken word recognition, lexical access.

Financement et collaboration

Cette thèse a été financée par une allocation de recherche du Ministère de l'Enseignement Supérieur et de la Recherche. Elle a été réalisée au sein du Laboratoire de Psychologie et de NeuroCognition (CNRS UMR 5105) et de l'Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement. Elle a été réalisée en collaboration avec Christophe Savariaux et Lionel Granjon du département Parole et Cognition du Gipsa-lab.

Remerciements

Je tiens tout d'abord à remercier mes directrices de thèse, pour avoir su m'encadrer tout au long de ces années. Merci à vous deux pour votre soutien, votre écoute, votre enthousiasme et votre énergie. Merci pour votre gestion humaine de la recherche. J'ai beaucoup appris à vos côtés. Sonia, je vous suis sincèrement reconnaissante de m'avoir permis de travailler sur ce sujet de recherche et de m'avoir laissée m'approprier ce projet qui, initialement, était le vôtre. Merci de m'avoir donné la chance de travailler avec vous, mais également de collaborer avec le Département Parole et Cognition du Gipsa-lab. Elsa, je te remercie d'avoir su me donner l'envie de travailler dans la recherche en psycholinguistique. Sache que depuis ton cours de licence, je suis fan de l'élision du schwa. Merci de m'avoir fait partager tes expériences passées, tu m'as évité de nombreux écueils. Merci également de t'être portée volontaire pour être la locutrice de mes manipes, même si tu ne te doutais pas forcément que cela impliquerait que des photos de toi soient disséminées tout au long de ce manuscrit.

Je remercie sincèrement, Juan Segui, Pascal Barone, Pierre Hallé, Uli Frauenfelder et Jean-Luc Schwartz pour avoir accepté d'évaluer mon travail. Je tiens à adresser un remerciement spécial à Jean-Luc, pour avoir répondu présent en tant que locuteur, directeur de l'EDISCE et finalement examinateur. Merci d'avoir successivement accepté de jouer ces trois rôles.

Je souhaite également remercier Denis Burnham, Cathi Best, Christine Kitamura et Chris Davis pour leur accueil chaleureux et extrêmement enrichissant au MARCS. Cathi, et Christine, merci pour votre gentillesse, cela a été un vrai plaisir de travailler avec vous, j'ai beaucoup appris.

J'aimerais adresser un grand merci à Christophe et Lionel du site Stendhal du Gipsa-lab qui m'ont beaucoup aidé dans ce travail. Chris, merci de ta patience, de tes compétences multiples et de ton sourire. Je te promets qu'à l'avenir, je ne dirai jamais plus que 0 dB correspond à une absence de bruit. Lionel, merci de m'avoir évité de nombreux déboires avec E-prime. Je te suis reconnaissante de m'avoir fait partager ton savoir technique et théorique, mais également d'être aussi têtu.

Pour continuer sur le thème Gipsa-lab, je tiens à remercier les doctorants du DPC, avec lesquels j'ai eu le plaisir de participer à l'organisation des RJCP et que je considère comme de véritables amis. Merci à Sandra (et mini-m), mais également à Amélie, Benj, Hien, Ros, Atef, pour m'avoir permis de partager cette aventure avec vous. Merci également à Anne V, Maria, Sylvia, Guillaume et à tous les autres membres du site Stendhal, pour l'accueil chaleureux dans vos locaux.

Ensuite, je voudrais exprimer ma reconnaissance à tous les membres du LPNC. L'ambiance a toujours été agréable et les échanges enrichissants, tant sur le plan professionnel que personnel. Un grand merci à Ronald Peereman pour son intérêt, ses précieux conseils et ses idées. Merci à Sylviane pour son soutien moral et matériel, merci à Richard et Stéphane pour les urgences statistiques et méthodologiques. J'aimerais décerner une mention spéciale à tous les jeunes chercheurs que j'ai côtoyé dans le bureau 222bis ou dans les locaux du BSHM: Mumu, Anne H2B, Anne T., Solène A., Jen B., Benoit M., Fleur, Alice, Yanica, Gaëtan, Seb C., Benjamin, (Jean) Sébastien, Marie L, Ben F., Nico, Cédric, Marcela, Lucie, Marie-Pierre. Merci à tous pour ces bons moments et pour avoir su gérer mes craquages. Je tiens aussi à remercier les relecteurs : Anne H2B (notre mère à tous), (Jean)

Sébastien, Solène A, Anne T, Anne V. Grâce à vous, ce manuscrit est moins truffé d'anglicismes tels que « les participants étaient instruits » ou « les auteurs ont reporté que ». Un merci tout particulier à Mumu, pour son « oursattitude », son amitié et son soutien dans cette période rédactionnelle. N'oublie pas que je serai toujours là si un DCMD te mets des bâtons dans les roues. Je tiens également à adresser mes remerciements à David, Gino et Julien D, pour leurs discussions et leurs conseils avisés. Merci aussi à Eric et à Nicole, pour leur soutien technique et moral.

Dans un tout autre registre, je tiens également à remercier toute la team de l'EVUG : merci au PAF Crew (maman Gâteau, Lélé, Marion), mais aussi à Bandou bandou, Monsieur P, Bik, Tom, M2R, Caneton, Freeze, Groyo, Kro, bubulle, Mika. Merci de m'avoir trimballée dans vos camions de spots en spots et d'avoir partagé le plaisir de la planche, du vent et du ti punch pendant toutes ces années. Merci pour votre grande amitié et pour tous ces bons moments. J'espère que beaucoup d'autres restent encore à venir. Je tiens à remercier au passage mon « home spot », le Monteynard, pour son vent thermique et son ambiance familiale.

J'aimerais également adresser une spéciale dédicace à mes colocs du Rif Talon, Jeff, Pierre et Rom-Rom pour leur « big brother attitude ». Jeff, merci pour ton écoute ton amitié sincère et ta gentillesse. Merci de m'avoir permis de traverser sans trop d'encombres cette année 2011, qui s'est avérée riche en rebondissements professionnels et personnels. Merci pour ces discussions philosophiques, ces échanges et ton soutien sans faille. Merci à Pierre pour tes pierrades et pour toutes ces soirées improvisées autour d'une bonne bouteille. Merci à Rom-Rom pour ta fraîcheur et ton incroyable énergie ainsi que pour ta spontanéité. Merci de ta capacité à me faire rire et à métamorphoser le jardin en l'espace d'une après-midi.

Merci à Chloé B : ton énergie et ton sourire m'ont permis de sortir de ma torpeur rédactionnelle, et de partager les affres de la relation à distance. Merci pour ta complicité et ton écoute.

J'aimerais également remercier Nadège, une Amie de longue date. Nad, merci pour ton amitié loyale et sincère, ton éternel optimisme. Merci d'avoir pris soin de moi depuis le lycée, je sais que je pourrais toujours compter sur toi.

J'aimerais aussi remercier mes parents ainsi que mes sœurs et mon frère. Merci d'être Fort et d'être toujours là, quoi qu'il arrive. Youssette, Bic, Armelle et Bérange, je vous remercie d'avoir toujours su veiller sur moi. Merci à mes parents de m'avoir insufflé la volonté d'aller jusqu'au bout des choses. Merci pour votre éducation, merci de m'avoir transmis vos valeurs. Je suis fière aujourd'hui de pouvoir dire qu'elles sont miennes.

Enfin, merci à Antoine pour ton amour, ton humour indécrottable, ta tendresse et ton incroyable patience. Merci de partager ma vie et d'être toujours présent quelles que soient les épreuves ou la distance.

Bonne lecture
Mathilde

Table des matières

Liste des abréviations

NB : Les abréviations en italique sont liées aux statistiques présentées dans ce manuscrit.

A Auditif, Auditive, Audio

AV Audiovisuel, Audiovisuelle

V Visuel, Visuel

ANOVA Analyse de variance (analysis of variance)

CV Consonne-Voyelle

CVC Consonne-Voyelle-Consonne

CVCV Consonne-Voyelle-Consonne-Voyelle

dB Décibels

η^2_p *Eta carré partiel (taille de l'effet)*

F Indice statistique suivant la loi de Fisher

GA Gyrus Angulaire

GSM Gyrus Supra Marginal

GTI Gyrus Temporal Inférieur

GTM Gyrus Temporal Moyen

GTS Gyrus Temporal Supérieur

ISI : Intervalle Inter Stimuli

I Intelligibilité

I_c Intelligibilité consonantique

IRMf Imagerie par Résonance Magnétique fonctionnelle

ISI Intervalle Inter Stimuli

M Moyenne

opm occurrences par million

p Probabilité associée à un indice statistique

r Coefficient de corrélation

RSB Rapport Signal Sur Bruit

STS Sillon Temporal Supérieur

t Indice statistique suivant la loi de Student

TMS Stimulation Magnétique Transcrânienne

VC Voyelle-Consonne

VOT Voiced Onset Time

Préambule

A la fin du XIX^{ème} siècle, l'établissement des premières lignes téléphoniques a incité la « Bell Telephone Company » aux Etats-Unis à conduire une étude visant à déterminer l'efficacité d'une communication orale uniquement basée sur la perception d'un signal acoustique (Argyle & Cook, 1976; cité par Locke, 1993). Près de 150 ans plus tard, un examen du succès grandissant des nouvelles technologies de la communication pourrait nous conduire au constat erroné que la perception de la parole est aujourd'hui devenue un évènement purement auditif (e.g., Massaro & Jesse, 2007; Rosenblum, 2005).

Or, 50 années de recherches dans des domaines aussi variés que la linguistique, la phonétique, la psycholinguistique, la neuropsychologie, la neurophysiologie etc. ont montré l'importance de considérer la parole comme un objet *multimodal* (e.g., Bernstein, Burnham, & Schwartz, 2002; Rosenblum, 2005, 2008). Plus spécifiquement, il est aujourd'hui communément admis que la perception de la parole est un évènement majoritairement *bimodal* ou *audiovisuel*. En effet, dans la vie quotidienne, excepté lorsque nous utilisons notre « smartphone » ou lorsque nous écoutons la radio, la plupart des situations de perception de la parole se déroulent face à face ou en modalité audiovisuelle (e.g., à la télévision).

Ainsi, pour une grande partie de ses interactions orales, l'être humain va non seulement percevoir une information auditive mais a également un certain nombre d'indices *visuels* à sa disposition. Ces derniers correspondent à la source de production du signal acoustique de parole et désignent ici les mouvements des différents articulateurs *visibles* du visage de notre interlocuteur (i.e., lèvres, mâchoire, dents, langue, etc.)[1].

Un des premiers bénéfices liés à la présence de cette information est très bien illustré par la proposition « Aspetta un attimo, mi metto gli occhiali che non ti sento bene [2]» (Franca Pugno, Rome, Italie, 1994).

En effet, cette courte phrase résume très bien la conclusion à laquelle sont arrivés Sumby & Pollack en 1954, dans leur étude pionnière « oral speech intelligibility may be appreciably improved in many practical situations by arrangement for supplementary visual

[1] Dans ce manuscrit, les termes « gestualité oro-faciale », « modalité visuelle », « signal visuel de parole », « information visuelle » et « parole visuelle » seront préférentiellement utilisés pour désigner cette source d'information.
[2] « Attends un moment, je mets mes lunettes parce que je ne t'entends pas bien »

observation of the speaker[3] » (p. 215). Ainsi, lorsque le signal acoustique est détérioré par du bruit environnant, nous savons que le fait de voir le visage de notre interlocuteur permet d'augmenter l'intelligibilité du signal acoustique de parole. Ce résultat, primordial pour l'étude de la parole, montre que voir les mouvements oro-faciaux d'un locuteur permet de mieux percevoir les *sons* de parole, lorsque le signal acoustique est détérioré par du bruit.

L'objectif de cette thèse consiste à examiner si nous sommes capables d'extraire du *sens* de cette information visuelle. En d'autres termes, le but de ce travail est d'évaluer la contribution spécifique de la gestualité oro-faciale verbale (i.e., non émotionnelle) au processus de reconnaissance de *mots*.

Pour cela, dans la première partie de ce manuscrit, nous effectuerons un état de l'art des recherches sur la parole visuelle et quels bénéfices l'être humain « tout venant » peut tirer de la perception de cette information lorsqu'il décode le langage oral (Chapitre 1). Nous examinerons ensuite les différents travaux ayant étudié le processus de reconnaissance de mots (i.e., l'accès au lexique) en modalité auditive seule (Chapitre 2).

Dans la seconde partie, nous présenterons nos propres recherches. Celles-ci viseront à examiner le rôle de la gestualité oro-faciale dans le processus d'accès au lexique, en modalité audiovisuelle (Chapitre 3) et visuelle seule (Chapitre 4) chez l'adulte. Après avoir effectué un état de l'art des travaux réalisés chez l'enfant, nous présenterons une étude développementale visant à observer l'évolution de l'utilisation des indices visuels avec l'âge (Chapitre 5).

C'est dans la discussion générale (Chapitre 6) que nous envisagerons comment différents modèles psycholinguistiques décrivant le processus de reconnaissance de mots en modalité auditive pourraient rendre compte de nos résultats en intégrant l'information visuelle dans leur architecture. Dans la partie perspectives, nous discuterons du(es) rôle(s) éventuel(s) que le signal visuel de parole pourrait jouer dans l'activation des représentations lexicales. Nous formulerons enfin différentes propositions visant à décrire comment certains modèles d'accès au lexique pourraient rendre compte de l'intégration audiovisuelle de la parole.

[3] « L'intelligibilité de la parole peut être sensiblement améliorée pour un grand nombre de situations pratiques lorsque l'on s'arrange pour que le locuteur soit également visible »

CHAPITRE 1. LA PERCEPTION DE LA PAROLE VISUELLE ET AUDIOVISUELLE

"Nothing in your perceptual world communicates so much information so quickly as a human face. From a face, you can rapidly determine an individual's identity, gender, emotional state, intentions, genetic health, reproductive potential, and even linguistic message (through lip-reading)"

(Rosenblum, 2010). p. 179.

1.1. INTRODUCTION

L'entrée des mots-clés « perception de la parole » sur le moteur de recherche Wikipédia donne lieu à la définition suivante : « La perception de la parole est le processus par lequel les humains sont capables d'interpréter et de comprendre les sons utilisés dans le langage » (http://fr.wikipedia.org/wiki/Perception_de_la_parole). Cette définition représente assez bien la doxa. Sans être fausse, celle-ci est néanmoins incomplète. En effet, la perception de tout objet se caractérise, par définition, comme un évènement *multimodal* ou *multisensoriel*. Des informations issues de nos cinq sens sont disponibles pour que notre cerveau puisse extraire différentes caractéristiques de ce stimulus extérieur. La perception du langage oral n'échappe pas à cette règle. En effet, comme l'exprime très bien Lawrence Rosenblum (Rosenblum, 2008) « Speech perception is inherently *multimodal*. Despite our intuitions of speech as something we hear, there is overwhelming evidence that the brain treats speech as something that we hear, see and even feel » (p. 405). Concevoir la parole comme un évènement purement auditif constitue donc une définition réductrice et incomplète de ce phénomène. Le système visuel représente également une source d'information exploitable par notre système perceptif pour décoder le langage oral[4].

Le cas des individus malentendants est une illustration de cette affirmation. En effet, ces derniers sont capables de percevoir la parole alors même que leur perception du signal acoustique est grandement détériorée, voir absente (surdité totale). Pour pallier à leur déficit auditif, ces derniers développent des méthodes compensatoires qui consistent notamment à extraire des informations du signal *visuel* de parole. Cette information est relative aux mouvements des différents articulateurs *visibles* du visage de notre interlocuteur (i.e., lèvres, mâchoire, dents, langue, etc.). Ainsi, l'examen de cette situation suggère que lorsque l'être humain souffre d'un déficit de traitement du signal acoustique, ce dernier est capable d'utiliser d'autres indices, issus de la modalité visuelle, pour pallier au manque d'information auditive (voir Bernstein, Demorest, & Tucker, 2000; Erber, 1974; Rouger et al., 2007; Strelnikov et al., 2009, parmi de nombreux travaux à ce propos). L'objectif de ce chapitre consiste à évaluer l'apport de l'information visuelle à la perception de la parole chez

[4] Notons que le toucher constitue également une source d'information exploitable pour la perception de la parole. La méthode TADOMA (Alcorn, 1932; cité par Reed et al., 1985) est une technique qui a été développée à l'École Perkins, au Massachusetts, afin de permettre à des élèves sourds et aveugles de percevoir et produire la parole. Cette méthode consiste à placer l'élève en situation de communication face à face avec son interlocuteur. Son pouce est posé sur les lèvres de son partenaire, l'index sur la joue, et les autres doigts sur le cou. Ainsi, l'élève peut sentir l'ensemble des aspects physiques tels que la variation de la pression de l'air dans les joues, des lèvres, la vibration des cordes vocales mais également les mouvements des lèvres et de la mâchoire. Cette méthode a montré que l'élève est capable de récupérer à travers le toucher des indices lui permettant de percevoir la parole, indices qu'il ne peut percevoir à travers l'audition et la vision qui sont déficitaires (voir Fowler & Dekle, 1991, pour une influence du toucher sur la perception des sons).

l'individu normo-entendant. Pour cela, nous présenterons tout d'abord les parties du visage et du conduit vocal qui transmettent une information exploitable par notre système visuel lors de la perception d'un signal de parole. Puis, nous examinerons quel type d'information peut être extrait de cette parole visuelle. Nous étudierons également plusieurs situations dans lesquelles voir le visage de son interlocuteur influence la perception du signal de parole chez l'adulte. Nous présenterons ensuite comment différents modèles envisagent la perception du langage oral en modalité audiovisuelle. Nous décrirons les mécanismes, le type de codage ainsi que le type de stockage postulés par chacun d'entre eux. Nous terminerons ce chapitre en présentant brièvement différentes structures cérébrales qui semblent spécifiquement impliquées dans le décodage de la parole en modalité audiovisuelle et visuelle seule.

1.2. LE SIGNAL VISUEL DE PAROLE

1.2.1. Les mouvements articulatoires visibles de l'appareil phonatoire de l'adulte

Avant de s'attarder sur l'aspect « perception » de la parole visible, nous allons brièvement présenter les différents articulateurs et résonateurs mis en jeu lors de la production de la parole visuelle en français et auxquels nous ferons référence tout au long de ce manuscrit. Ainsi, le langage oral est avant tout un signal physique produit par la mise en action séquentielle et/ou simultanée de différents articulateurs (ou organes) et résonateurs (ou cavités, cf. Figure 1).

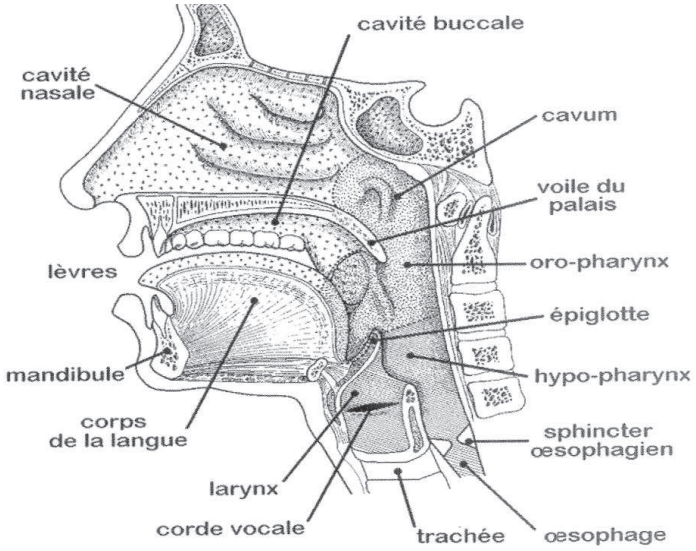


Figure 1. *Le tractus vocal : principaux repères anatomiques pour une description articulatoire de la production de la parole. (Extrait de Marchal, 2011).*

L'air, expulsé par les poumons, produit le souffle qui emprunte la trachée puis permet aux cordes vocales de rentrer en vibration. Il passe ensuite à travers différentes cavités telles que le larynx, le pharynx, la cavité buccale et parfois la cavité nasale. Les mouvements des lèvres, de la mâchoire, de la langue, du voile du palais, etc. permettent de changer la forme du tractus vocal et de produire ainsi des sons différents. Néanmoins, bien que la conséquence de la mise en action de l'ensemble de ces organes soit audible, seuls certains de ces articulateurs sont visibles en situation de communication face à face. Leur fermeture ou le fait que la langue touche une partie supérieure du conduit vocal permet de créer un obstacle temporaire à la diffusion de l'onde sonore dans l'air. L'air va s'accumuler temporairement à l'endroit où se trouve cet obstacle. Le relâchement de cette occlusion va engendrer une « plosion » acoustique dans le signal sonore, permettant de produire les consonnes dites « plosives » ou « occlusives » (e.g., /p/, /b/, /m/, /t/, /d/, /n/, /k/, /g/, /ŋ/). Lorsque les lèvres ou la langue n'opposent pas une fermeture complète mais réduisent seulement la taille du conduit vocal, des turbulences vont être créées au passage de l'air à cet endroit. Ces turbulences vont être à l'origine d'un phénomène de « friction » dans le signal acoustique. Le mouvement articulatoire de constriction à l'origine de ce phénomène permet de produire des consonnes dites « fricatives » ou « constrictives » (e.g., /f/, /v/, /s/, /z/, /ʃ/, /ʒ/). Notons que ces deux modes d'articulation (i.e., constrictif et occlusif) sont surtout saillants dans le signal visuel lorsque les consonnes sont réalisées au niveau des lèvres (i.e., /f/ et /p/) ou engendrent un mouvement spécifique des lèvres (i.e., /ʃ/, /g/) plutôt que lorsqu'ils résultent d'une occlusion ou d'une constriction liée au positionnement de la langue derrière les dents (i.e., /t/ et /s/). Le lieu (ou la place) d'articulation est une caractéristique du signal visuel de parole qui est particulièrement saillante. Celle-ci permet par exemple de générer différentes consonnes à l'intérieur des catégories « occlusives » (e.g., /b/, vs. /d/ vs. /g/) et « constrictives » (e.g., /v/ vs. /z/ vs. /ʒ/). Le mouvement haut bas de la partie inférieure de la mâchoire constitue également un mouvement articulatoire facilement perceptible dans la parole visuelle : son degré d'aperture (ou d'ouverture) permet de modifier la taille de la cavité buccale par laquelle va circuler l'onde sonore engendrant ainsi la production de voyelles différentes (e.g., /a/ vs. /o/, /y/, /u/). Enfin, la dimension arrondissement-étirement des lèvres est également une caractéristique hautement visible, qui permet aussi de modifier l'écoulement de l'air dans le tractus vocal et modifie de ce fait le signal acoustique de certaines voyelles (e.g., /y/ vs /i/). Notons ici que la nasalisation, résultant de l'abaissement du voile du palais et permettant de faire entrer la cavité nasale en résonance (e.g., /m/ vs. /b/ ; /a/ vs /ã/), le voisement, engendré par la vibration plus ou moins longue des cordes

vocales (e.g., /b/ vs. /p/) ainsi que la position de la langue lorsque le degré d'aperture de la mâchoire est faible (e.g., /y/ vs. /u/) sont des mouvements articulatoires difficilement perceptibles dans le signal visuel de parole.

Ainsi, d'après cette description, nous pouvons conclure que la production du langage oral est une activité complexe dont l'origine est *articulatoire*. En d'autres termes, cela signifie que la production des signaux acoustiques et visuels de parole résulte de l'action combinée de différents organes et résonateurs qui vont venir modifier les mouvements visibles de la face ainsi que l'onde sonore issue des cordes vocales (cf. Figure 2).

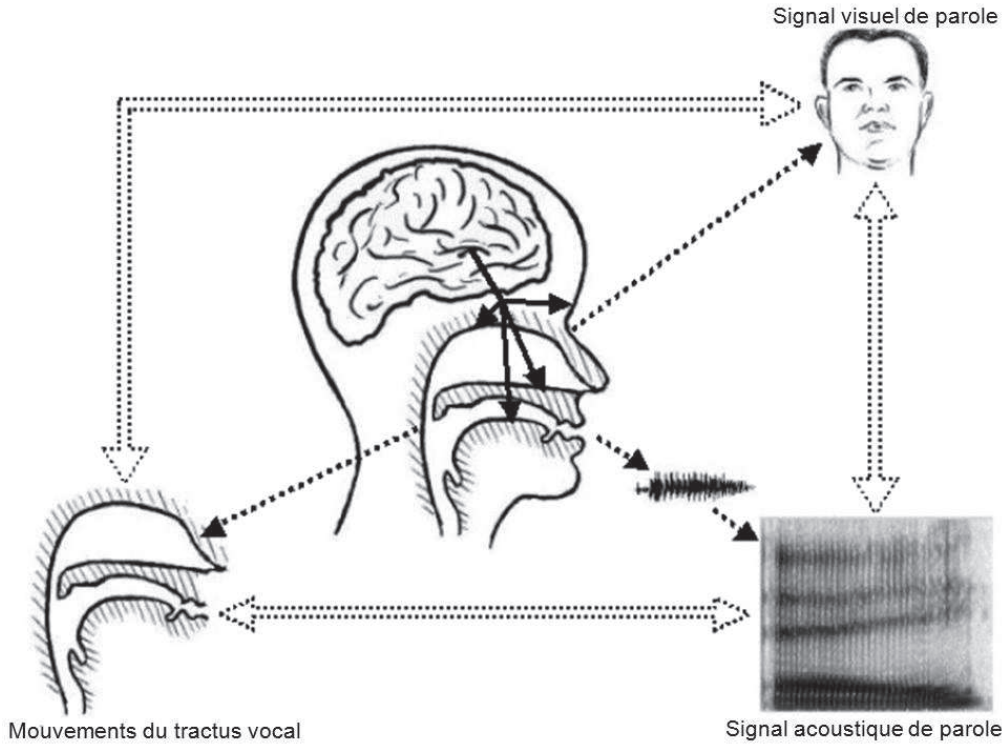


Figure 2. *Relation entre les mouvements oro-faciaux visibles (signal visuel de parole), les mouvements des articulateurs du tractus vocal et le signal acoustique de parole. (Adapté de Jiang, 2003).*

Dans cette partie, nous avons donc qualifié les principaux mouvements articulatoires présents dans le *signal* visuel de parole. Nous avons souligné que certains d'entre eux sont, particulièrement visibles sur le plan *perceptif* (e.g., la place d'articulation), alors que d'autres le sont moins (e.g., mode d'articulation). L'objectif de la section suivante est d'étudier quels articulateurs sont exploités par l'individu pour en extraire de l'information linguistique. Nous présenterons ensuite les différentes situations dans lesquelles un interlocuteur normo-entendant bénéficie de la présence du signal visuel de parole.

1.2.2. Apport des différentes parties du visage du locuteur à la perception du signal visuel de parole

Certaines études se sont intéressés à évaluer la part respective des différents éléments du visage dans la transmission d'une information exploitable dans le signal visuel de parole (e.g., Badin, Tarabalka, Elisei, & Bailly, 2010; Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996, cités par Schwartz, 2011 ; Cvejic, Kim, & Davis, 2010; Guiard-Marigny, Tsingos, Adjoudani, Benoît, & Gascuel, 1996; McGrath, 1985, cité par Summerfield, 1991; Munhall & Vatikiotis-bateson, 1998; Preminger, Lin, Payen, & Levitt, 1998; Summerfield, 1991; Thomas & Jordan, 2004; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998, voir Thomas & Jordan, 2004, pour un état de l'art relativement récent et détaillé à ce sujet).

Conformément ce que l'on pourrait penser, voir les lèvres du visage de son interlocuteur permet de transmettre une part importante de la composante visuelle du langage oral. Benoît et al. (1996, cités par Schwartz, 2011) ont par exemple montré que l'apport informationnel transmis par la présentation des lèvres seules représentait les deux tiers des informations véhiculées par la présentation d'un visage entier. D'autres travaux ont également montré que les dents (McGrath, 1985; cité par Summerfield, 1991; Thomas & Jordan, 2004), la mâchoire (e.g., Guiard-Marigny, et al., 1996; Vatikiotis-Bateson, et al., 1998) et la langue (Badin, et al., 2010), constituent également une source d'information importante pour percevoir le signal visuel de parole. Soulignons que ce dernier travail met en évidence notre capacité à décoder plusieurs contrastes phonétiques lorsque les mouvements de la langue sont rendus entièrement visibles pour l'interlocuteur, alors que dans une situation de communication face à face, ces mêmes indices ne sont que partiellement perceptibles dans le signal visuel de parole. Plus précisément, cette étude indique qu'un individu normo-entendant est capable de décoder un mouvement articulatoire habituellement non visible (e.g., position arrière/avant de la langue) lui permettant de distinguer deux phonèmes (e.g., /y/ vs. /u/), alors même qu'il ne dispose que de peu ou même d'aucune expérience perceptive visible avec ce contraste (e.g., à cause de l'arrondissement important des lèvres) dans les situations de communication face à face de la vie quotidienne.

Certains travaux ont également montré que les performances en lecture labiale[5] étaient meilleures lorsque le visage entier du locuteur était présenté par rapport à une situation où seuls les mouvements de la zone orale (i.e., correspondant à la présentation des

[5] Notons ici que ce terme sera employé tout au long de ce manuscrit mais qu'il ne désigne pas seulement la capacité d'un individu à extraire de l'information linguistique à partir des mouvements des lèvres, mais également à partir de l'ensemble des articulateurs de son interlocuteur.

lèvres, de la langue, des dents et de la mâchoire) sont visibles (e.g., Thomas & Jordan, 2004). Ce bénéfice supplémentaire suggère que même si les mouvements de la partie buccale constituent un moyen efficace de véhiculer le signal visuel de parole, d'autres informations situées autour de cette zone sont également exploitables (voir Benoît, et al., 1996, pour des résultats similaires). En effet, il a été montré que les joues (Preminger, et al., 1998) ainsi que le haut de la tête (e.g., Cvejic, et al., 2010; Munhall & Vatikiotis-bateson, 1998) transmettent aussi des indices relatifs au langage parlé. Par exemple, de récents travaux (Cvejic, et al., 2010) indiquent que des indices prosodiques, faisant référence aux phénomènes d'accentuation de l'intonation de la langue, sont véhiculés par les mouvements de la tête du locuteur.

Une étude effectuée par Vatikiokis-Bateson et al. (1998) a permis d'évaluer en temps réel les parties du visage qui étaient préférentiellement traitées par des participants en modalité audiovisuelle, lorsque le signal acoustique de parole est détérioré par du bruit. En utilisant une méthode d' « eye-tracking[6] », ces auteurs ont mis en évidence que la bouche mais également les yeux du locuteur étaient préférentiellement regardés par rapport au reste du visage. Ils ont également observé que le temps de fixation sur la zone buccale représentait 35 % du temps de fixations total sans bruit mais augmentait significativement (i.e., jusqu'à 55 % du temps de fixations total) avec l'augmentation de la dégradation de l'information auditive. Parallèlement, le temps de fixation total sur les yeux du locuteur diminuait avec l'augmentation de la proportion de fixations sur la bouche. Ces résultats indiquent donc que lorsqu'un individu traite l'information visuelle, il a majoritairement recours aux mouvements des articulateurs situés dans la zone buccale (i.e., des lèvres, de la mâchoire, de la langue et des dents) pour augmenter l'intelligibilité du signal acoustique de parole.

En résumé, l'examen de ces différents travaux permet de montrer que voir le visage de son interlocuteur dans sa globalité permet d'apporter plus d'informations en comparaison avec une situation où seuls les mouvements des articulateurs situés dans la partie inférieure de la face sont visibles (Thomas & Jordan, 2004). Ainsi bien que la majorité de l'information visuelle soit véhiculée par la zone buccale (e.g., Benoît, et al., 1996; McGrath, 1985, cité par Summerfield, 1991 ; Thomas & Jordan, 2004), la totalité des mouvements du visage et de la tête du locuteur fournissent aussi des indices utiles à la perception de la parole (e.g., Cvejic, et al., 2010; Munhall & Vatikiotis-bateson, 1998; Preminger, et al., 1998). L'objectif de la partie suivante consiste à caractériser le type d'information linguistique véhiculée par le signal visuel seul de parole lorsque l'intégralité du visage en mouvement du locuteur est visible.

[6] Enregistrement des mouvements oculaires

1.2.3. Nature visémique de l'information visuelle

Une des raisons pour lesquelles tout un pan de la psycholinguistique et de la phonétique étudie la perception de la parole en modalité auditive seule vient du fait que le signal acoustique constitue, au moins dans certaines situations (i.e., au téléphone, à la radio) une information suffisante pour décoder un mot isolé produit par un locuteur. Or, il n'est pas possible d'en dire autant pour l'information visuelle. En effet, excepté pour des mots très fréquents (e.g., MacLeod & Summerfield, 1987, cité dans Summerfield, 1991) ou très prédictibles vis-à-vis de la situation de communication, les mots ne sont que très difficilement identifiables sur la seule base du geste articulatoire de parole. A titre d'exemple, il a été évalué que cette information permet de discriminer 40 à 60 % des phonèmes d'une langue et de 10 à 20 % des mots (e.g., Schwartz, 2011). Cette difficulté d'identification vient du fait que certaines caractéristiques acoustiques (e.g., le voisement, la nasalisation) permettant de distinguer deux phonèmes ou sons de parole (e.g., /b/ vs. /p/ ; /b/ vs. /m/, respectivement) correspondent dans le signal visuel à la mise en action d'articulateurs non visibles du conduit vocal (e.g., vibration des cordes vocales situées dans le larynx ; abaissement du voile du palais, cf. section 1.2.1). Ainsi, certains phonèmes ne peuvent pas être discriminés consciemment sur la seule base de l'information visuelle (e.g., Fisher, 1968; Summerfield, 1987, 1991), ce qui rend la distinction entre certains mots (e.g., « pain » vs. « bain ») très difficile voire impossible en modalité visuelle seule (voir e.g., Mattys, Bernstein, & Auer, 2002, pour plus de détails à ce sujet). Notons que dans cette section, nous allons présenter différents travaux qui se sont intéressées à classifier les phonèmes selon leur intelligibilité en modalité visuelle seule en anglais et en français. Nous nous intéresserons aux études ayant évalué la possibilité de discrimination de *mots* en modalité visuelle seule (e.g., Mattys, et al., 2002) dans la section 4.1.1 du Chapitre 4 de ce manuscrit.

Ainsi, plusieurs études basées sur des matrices de confusion ont cherché à établir quels phonèmes étaient confondus ou distingués en modalité visuelle seule en anglais (e.g., Fisher, 1968; Jiang, 2003; Walden, Prosek, Montgomery, Scherr, & Jones, 1977, cité dans Summerfield, 1991, voir Jiang, 2003 pour une revue), mais aussi en français (e.g., Gentil, 1981; Jutras, Gagné, Picard, & Roy, 1998; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998, voir Aboutabit, 2007 et Cathiard, 1994 pour des revues). Summerfield (1991) exprime d'ailleurs très bien ce qu'intuitivement nous pourrions penser à ce sujet « If we attempt to

answer the question "How well can the phonetic module lipread consonants and vowels ?" we might conclude "Not very well"[7] » (p. 118).

Une des premières recherches a donné lieu à l'établissement de différentes catégories appelées visèmes (Fisher, 1968). Ce terme vient de la contraction du mot « visuel » et « phonème ». Selon Fisher (1968), la notion de visème renvoie à toute unité minimale de la parole qui est perçue visuellement distincte d'une autre unité. Pour les phonèmes consonantiques, une classification récente et relativement complète des différents visèmes de la langue anglaise a été effectuée durant les travaux de thèse de Jiang (Jiang, 2003). Pour cela, il a utilisé 69 syllabes de type CV correspondant à la combinaison de trois voyelles et 23 consonnes. Son travail a permis de mettre en évidence une très bonne intelligibilité pour les phonèmes /w/, /h/ and /f/, alors que les phonèmes /r/, /n/ et /g/ semblent être beaucoup plus facilement confondus. Ainsi, il a pu dégager six catégories de visèmes {w} {p,b,m} {r,f,v} {θ, ð} {l,n,k,g,h,j} {t,d,s,z,ʃ,ʒ,tʃ,dʒ}, (voir Walden, et al., 1977, cité par Summerfield, 1991, pour une classification en neuf catégories visémiques). Pour les phonèmes consonantiques du français, une classification a été effectuée par Gentil (Gentil, 1981). Son travail a permis de définir différentes catégories en fonction de la position (initiale ou finale) de chacune des 16 consonnes étudiées. Ainsi, il a pu mettre en évidence quatre visèmes pour des consonnes situées à l'initiale : {p,b,m} {f,v} {ʃ,ʒ} et {s,z,t,d,n,k,g,ŋ,ʁ}. En position finale, quatre groupes distincts supplémentaires semblent pouvoir être dégagés de la dernière catégorie : {s,z} {t,d,n}, {k,g,ŋ} {ʁ}. Les phonèmes consonantiques les plus facilement distingués en position initiale correspondent à ceux qui sont articulés à l'avant du conduit vocal {p,b,m} {f,v} (i.e., au niveau des lèvres) et ceux qui disposent d'une caractéristique articulatoire hautement visible {ʃ,ʒ} (i.e., mouvement de protrusion des lèvres). Inversement, les phonèmes consonantiques les plus facilement confondus étaient articulés au niveau ou derrière les dents, c'est-à-dire plus en arrière du conduit vocal {s,z,t,d,n,k,g,ŋ,ʁ}. Ce constat renvoie à la notion de saillance perceptive (e.g., Summerfield, 1987) ; elle sera détaillée dans la section 1.3.1.3 de ce chapitre.

Pour les voyelles, aucune classification aussi claire n'a pu être effectuée jusqu'à présent. Ainsi, il semblerait que l'on ne puisse pas constituer, comme pour les consonnes, de groupes nets de visèmes vocaliques (e.g., Heider & Heider, 1940, cité dans Cathiard, 1994). En français, l'étude de Gentil (1981) a mis en évidence que le /a/, le /i/, le /ɛ/ et le /u/ étaient facilement identifiables en modalité visuelle seule. Le /y/ était très souvent confondu avec le /u/. A propos de ce travail, Cathiard (Cathiard, 1994) remarque que les confusions

[7] Si nous essayons de répondre à la question "Comment notre module phonétique peut-il lire sur les lèvres les voyelles et les consonnes ?" nous pourrions conclure "Pas très bien" »

avaient exclusivement lieu à l'intérieur du groupe des voyelles arrondies (i.e., {œ,o,ɔ,õ,u,y,ø}) où à l'intérieur du groupe de voyelles non arrondies (i.e., {e,ɛ,i,ɛ̃,ã,a}). Ce résultat indique donc que la dimension (ou trait) d'arrondissement-étirement labial constitue une information visuelle pour l'identification des voyelles (Cathiard, 1994, voir section 1.3.3 pour plus de détails à ce sujet).

Un autre trait a également été envisagé pour la discrimination des voyelles du français : celui de la hauteur (ou séparation inter-labiale), correspondant au degré d'aperture (ou d'ouverture) du conduit vocal (et de la mâchoire) lors de la production d'une voyelle (e.g., Robert-Ribes, et al., 1998). La Figure 3, représente les résultats de Robert-Ribes et al. (1998) concernant les mesures articulatoires effectuées sur différentes voyelles du français (i.e., /a/, /ɛ/, /i/, /e/, /œ/, /ɔ/, /y/, /ø/, /u/, /o/).

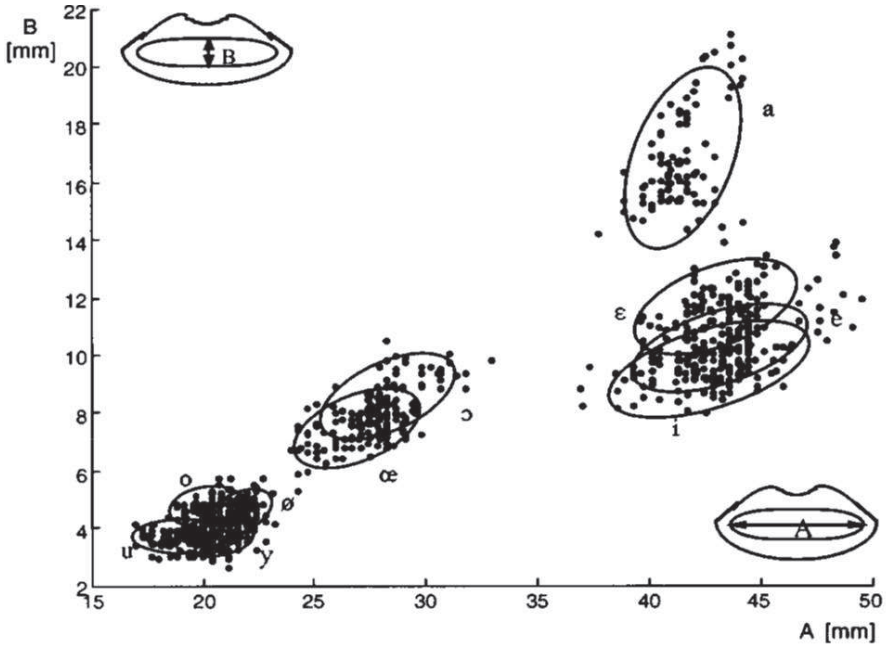


Figure 3. *Représentations de 10 voyelles orales du français en fonction des paramètres géométriques de hauteur (B) et d'arrondissement-étirement des lèvres (A). (Extrait de Robert-Ribes, et al., 1998).*

Ces auteurs ont ensuite effectué un test perceptif en modalité visuelle seule sur une partie de ces voyelles /a/, /e/, /i/, /o/, /u/, /y/, /ø/. Il ont mis en évidence les trois visèmes suivants {a}, {e,i}, {o,u,y,ø}, suggérant que les dimensions d'arrondissement-étirement (e.g., distinguant un /y/ d'un /i/) mais également de hauteur (e.g., distinguant un /a/ d'un /i/ et d'un /y/, cf. Figure 3) sont des traits pertinents pour la discrimination perceptive des voyelles en modalité visuelle seule (voir Jackson, Montgomery, & Binnie, 1976, pour des résultats similaires en anglais).

En conséquence, l'ensemble de ces études indique que même si nos différents articulateurs ne sont pas toujours visibles lors de la perception du signal de parole, l'être humain adulte dispose d'une capacité de traitement relativement fine des informations visuelles dont il dispose. Summerfield (1991) conclut à ce propos « an alternative answer to the question "How well can the phonetic module lipread consonants and vowels ?" could reasonably be given as "Quite well, given the limited evidence available to it"[8] » (p. 119).

Notons ici que le contexte vocalique/consonantique dans lequel un phonème est inséré influence sa perception en en modalité visuelle seule (e.g., Benguerel & Pichora-Fuller, 1982; Benoît, Mohamadi, & Kandel, 1994; Gentil, 1981) ; (voir Aboutabit, 2007 et Cathiard, 1994 pour des revues). Ce phénomène, appelé coarticulation, renvoie au fait qu'un phonème n'est pas articulé de la même manière en fonction des phonèmes adjacents. Dans la langue française, Benoît et al (1994) ont examiné l'identification de certaines consonnes (i.e., /b, v, ʁ, l, z, ʒ/) et de trois voyelles se situant aux extrémités des continua formés par les dimensions d'arrondissement-étirement et du degré d'aperture sur le plan articulatoire (i.e., /i/, /a/ et /y/ cf. Figure 3). L'ensemble des combinaisons entre ces consonnes et ces voyelles était présenté dans des séquences non significatives de type VCVCVz (e.g., /ababaz/ ; /ylylyz/, etc.). Leurs résultats ont montré qu'en modalité visuelle seule, l'intelligibilité des consonnes en fonction des différents contextes vocaliques utilisés (notée Ic) était ordonnée comme suit : $Ic_a > Ic_i > Ic_y$. Globalement, cela indique que la perception d'un phonème consonantique en modalité visuelle seule n'est pas influencée de la même manière en fonction des caractéristiques articulatoires des phonèmes vocaliques qui lui sont adjacents (voir Gentil, 1981; Massaro, Cohen, & Gesi, 1993, pour des résultats similaires en français et en anglais, respectivement). Cela vient du fait que la production d'un phonème consonantique va être plus ou moins contrainte (et donc plus ou moins visible) en fonction des caractéristiques articulatoires des phonèmes vocaliques adjacents. Par exemple, les réalisations articulatoires du /i/ dans /ibibiz/ et du /y/ dans /ybybyz/ impliquent une configuration labiale spécifique et contraignante (i.e., un étirement ou une protrusion des lèvres importante, respectivement), alors que la production d'un /a/ dans /ababaz/ peut être réalisée avec un degré d'aperture plus variable. Ainsi, une configuration articulatoire contraignante engendre, du fait de la coarticulation, une modification du mouvement d'occlusion pour le /b/ importante lorsque celui-ci est présenté dans le contexte d'un /i/ ou

[8] Par conséquent, une réponse alternative à la question « Comment notre module phonétique peut-il lire sur les lèvres les voyelles et les consonnes ? » peut raisonnablement être « Plutôt bien, étant donné la quantité limitée d'information disponible pour cela »

d'un /y/, le rendant plus facilement identifiable dans le contexte d'un /a/ (Benoît, et al., 1994).

Suite à l'examen de ces différentes études, nous pouvons conclure que du fait de la mise en action de certains articulateurs soit audible mais invisible (e.g., vibration des cordes vocales induisant la perception du voisement), l'information visuelle issue du signal de parole ne permet pas d'identifier l'ensemble des contrastes phonétiques définissant les phonèmes d'une langue. Néanmoins, les études montrent que nous sommes capables d'identifier en modalité visuelle seule une grande partie des traits articulatoires visibles (e.g., Gentil, 1981; Jiang, 2003; Robert-Ribes, et al., 1998; Summerfield, 1987). Cette capacité est cependant modulée en fonction de la place des phonèmes à identifier dans le mot (Gentil, 1981) ainsi que par la présence de phonèmes adjacents (e.g., Benoît, et al., 1994). L'objectif de la partie suivante est d'examiner les diverses situations dans lesquelles notre capacité d'extraction d'information du signal visuel de parole nous aide à percevoir le langage oral.

1.3. APPORT DE L'INFORMATION VISUELLE A LA PERCEPTION DE LA PAROLE

1.3.1. Perception audiovisuelle de parole en présence de bruit dans le signal acoustique

1.3.1.1. Travaux princeps

Dans la vie quotidienne, la perception de la parole s'effectue rarement dans un environnement complètement silencieux. Généralement, les lieux où se déroulent nos interactions langagières présentent des caractéristiques physiques (e.g., réverbération de l'onde sonore sur les parois d'une pièce), logistiques (e.g., bruit lié au fonctionnement d'un matériel électrique), naturelles (e.g., bruit du vent) et sociales (e.g., présence de signaux de parole concurrents) qui détériorent la composante auditive du langage oral. Plusieurs études ont montré que le fait de « tendre l'oreille » nous permettait de pallier à ce déficit informationnel et ainsi de mieux percevoir le discours de notre interlocuteur. Ce phénomène, appelé effet « Cocktail Party » désigne cette capacité selon laquelle notre système auditif parvient à sélectionner une source sonore pertinente dans un environnement bruité tout en traitant le reste des informations acoustiques (Cherry, 1953). Cependant, lorsque plusieurs sources sonores sont actives, des effets de masquage peuvent tromper notre perception et nous amener à confondre certaines d'entre elles. Notre système perceptif se sert alors des informations véhiculées par les autres sens (e.g., la vision) pour augmenter l'intelligibilité d'une source par rapport à une autre. Ainsi, en situation de perception audiovisuelle de la

parole (i.e., lorsque deux interlocuteurs communiquent face à face) nous parvenons à comprendre notre interlocuteur non seulement en concentrant notre attention auditive sur un seul flux de parole (effet « cocktail party ») mais aussi en traitant les informations véhiculées par la vision, c'est-à-dire la gestualité oro-faciale de notre partenaire. En effet, il est aujourd'hui communément admis que l'individu tout venant est capable de traiter l'information visuelle de parole transmise par le visage de son interlocuteur pour augmenter l'intelligibilité du signal acoustique, lorsque ce dernier est détérioré par du bruit (e.g., Benoît, et al., 1994; Binnie, Montgomery, & Jackson, 1974; Erber, 1969; Gagné, Rochette, & Charest, 2002; MacLeod & Summerfield, 1987; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Schwartz, Berthommier, & Savariaux, 2004; Sumby & Pollack, 1954; Summerfield, 1987).

A notre connaissance, les premiers travaux ayant mis en évidence de phénomène sont ceux de Sumby & Pollack (1954). Dans cette étude, les auteurs ont utilisé des listes de mots bisyllabiques présentés en modalité auditive (i.e., seule l'information auditive du signal de parole était disponible) ou audiovisuelle (i.e., l'information auditive et celle issue des mouvements du visage du locuteur étaient disponibles) à 129 participants de langue maternelle anglaise, ne disposant d'aucun entraînement spécifique à la lecture labiale. Le signal acoustique des stimuli était détérioré artificiellement à différents Rapport Signal sur Bruit (RSB). Précisons que le RSB, exprimé en Décibels, désigne le rapport entre la puissance du signal auditif et la puissance des bruits parasites comme par exemple les bruits de fond. Cette mesure permet d'évaluer le niveau de bruit dans lequel est inséré un signal acoustique. Ainsi, plus ce RSB est grand, plus le bruit est faible et plus un signal est perceptible. Inversement, plus un RSB est faible, plus le bruit est élevé et moins le signal est perceptible. Les résultats de l'étude de Sumby et Pollack sont reproduits dans la Figure 4a. L'analyse de leurs données montre clairement de meilleurs scores d'identification de mots en modalité audiovisuelle qu'en modalité auditive seule. De plus, l'examen de leurs performances montre que le bénéfice lié à la présence de la modalité visuelle augmente avec la détérioration du signal acoustique (i.e., avec la diminution du RSB). Ces résultats suggèrent qu'un individu adulte normo-entendant est capable de traiter l'information visuelle afin de mieux identifier un signal de parole détérioré par du bruit. Plusieurs autres études (e.g., Benoît, et al., 1994; Binnie, et al., 1974; Erber, 1969, cf. Figure 4b, c et d, respectivement) ont retrouvé ce même pattern de données, indiquant ainsi que décoder le signal visuel de parole augmente l'intelligibilité du langage oral présenté dans du bruit, lorsqu'une situation de perception audiovisuelle est comparée à une condition de perception auditive.

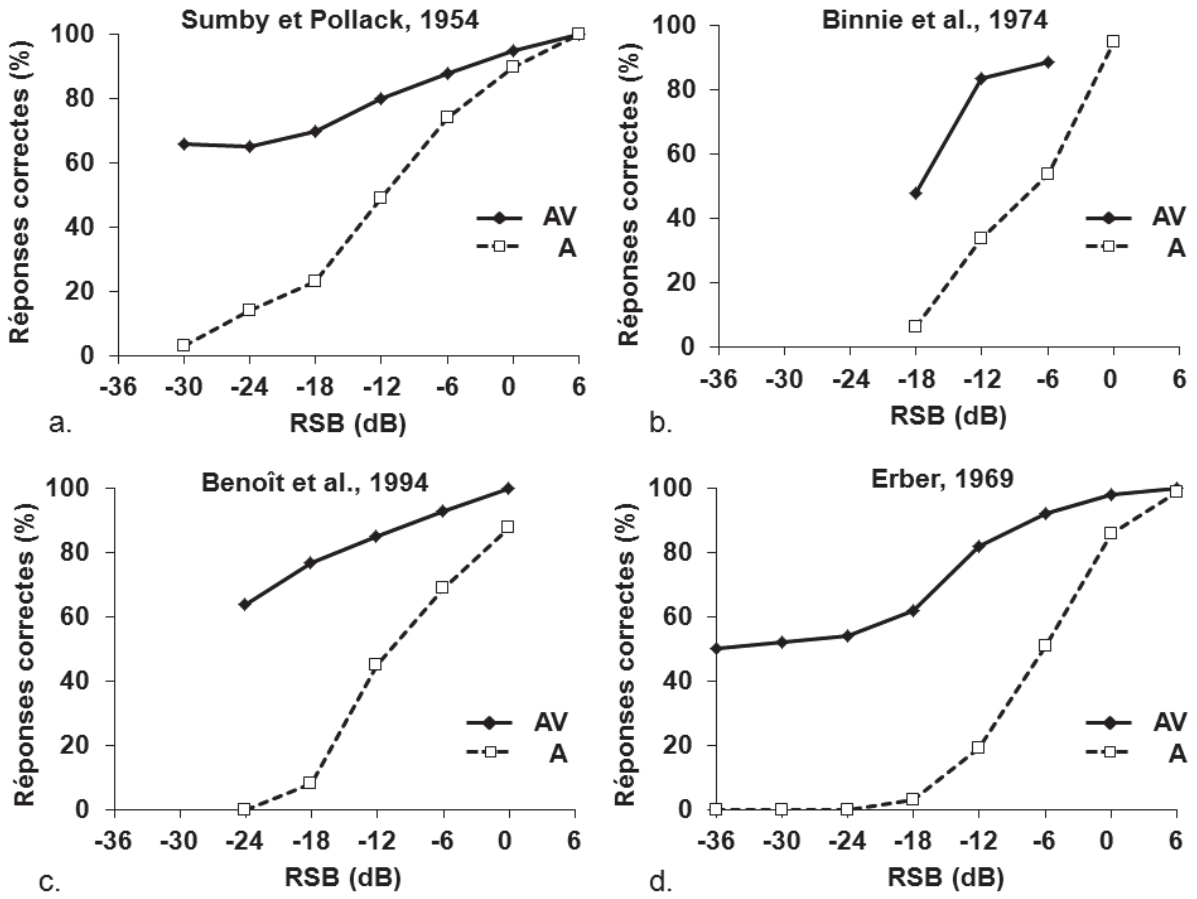


Figure 4. *Reproduction des résultats issus des études de (a) Sumby et Pollack, (1954), (b) Benoît et al. (1994), (c) Binnie et al. (1974) et (d) Erber (1969). Pour chaque graphique, la proportion de réponses correctes est exprimée en pourcentage (%) en fonction de la modalité de présentation des stimuli (A : Auditive vs. AV : Audiovisuelle) et du RSB (exprimé en dB). (Adapté de Schwartz, 2011).*

1.3.1.2. Quantification de l'apport de l'information visuelle

D'autres travaux ont également cherché à quantifier l'apport de la modalité visuelle en fonction des différents degrés de détérioration du signal acoustique (e.g., MacLeod & Summerfield, 1987; Ross, et al., 2007; Sumby & Pollack, 1954). Sumby et Pollack (1954) ont ainsi montré que dans leur étude, le fait de que percevoir de la parole en modalité audiovisuelle était équivalent à améliorer le rapport signal sur bruit de 15 dB en modalité auditive seule. Avec la même mesure, MacLeod et Summerfield (1987) ont mis en évidence que la présence de l'information visuelle correspondait à une amélioration du RSB de 11 dB. Dans une étude relativement récente Ross et al. (2007) ont proposé une tâche d'identification de mots monosyllabiques isolés, en modalité auditive, audiovisuelle ou visuelle seule. Pour l'ensemble des conditions, différents RSB étaient utilisés. La Figure 5 représente leurs résultats.

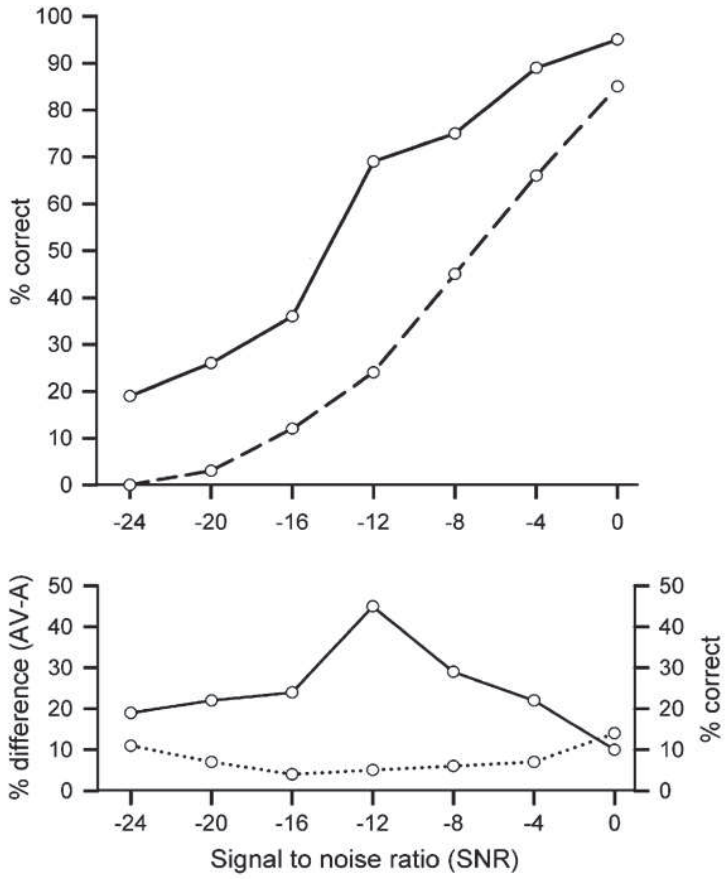


Figure 5. *Le graphique du haut correspond au pourcentage de mots correctement identifiés en fonction de la modalité de présentation des stimuli (Auditive : courbe en pointillés vs. Audiovisuelle : courbe continue) et du RSB (« SNR », en dB). Le graphique du bas correspond à la différence de mots correctement identifiés en modalité audiovisuelle et auditive. La courbe en pointillés correspond au pourcentage de mots correctement identifiés en modalité visuelle seule. (Extrait de Ross, et al., 2007).*

L'examen de leurs données (cf. graphique du bas) montre tout d'abord que comme pour les études précédentes, le bénéfice lié à la présence de l'information visuelle (noté «% difference AV - A ») augmente avec la diminution du RSB (noté « SNR ») pour des valeurs comprises entre 0 et -12 dB. La taille de ce gain atteint un pic à -12 dB (i.e., $M = 45$ %). La taille du gain commence à diminuer lorsque le RSB devient plus faible (pour des valeurs comprises entre -12 et -24 dB). Elle devient même beaucoup moins importante à -24 dB (i.e., autour de $M = 7$ %), lorsque l'information auditive seule n'est plus du tout intelligible (i.e., pourcentage d'identification correctes situé à autour de 0 % en modalité auditive seule). Ainsi, cette étude met en évidence que le gain d'intelligibilité lié à la présence de l'information visuelle augmente lorsque le signal acoustique commence à être détérioré ; il est maximal pour des RSB *intermédiaires* (i.e., à -12 dB dans cette étude), lorsque le signal acoustique est encore relativement intelligible ($M = 23$ % d'identifications correctes en modalité auditive seule, cf. graphique du haut). Cette étude remet donc en question la conclusion de Binnie et al. (1974) selon laquelle les bénéfices liés à la présence de

l'information visuelle augmentent plus le RSB est faible et moins l'information auditive est perceptible « greatest visual complement occurring at poorer SNRs[9] » (p. 629). Toujours dans l'étude de Ross et al. (2007), les résultats obtenus pour la condition visuelle seule (cf. graphique du bas, courbe en pointillés), montrent que les scores en lecture labiale étaient relativement identiques quel que soit le RSB. Les performances en modalité visuelle seule étaient relativement basses (i.e., $M = 9$ % tous RSB confondus). Ces performances étaient corrélées positivement avec le gain d'intelligibilité observé en modalité audiovisuelle par rapport à la modalité visuelle seule, mais seulement pour les RSB les plus faibles (i.e., à -20 et -24 dB). A -24 dB, les scores d'identification en modalité audiovisuelle étaient à $M = 19$ %, alors qu'aucun mot n'était identifié en modalité auditive seule (i.e., $M = 0$ %, cf. graphique du haut). Les auteurs font remarquer que même si les performances en modalité visuelle seule étaient corrélées avec le gain d'intelligibilité observé en modalité audiovisuelle, elles n'expliquent que 8-9 % de ce bénéfice. Ce résultat suggère donc que l'information linguistique extraite par un individu en modalité audiovisuelle est supérieure à la somme des indices que nous sommes capables de percevoir en situations unimodales, c'est-à-dire en modalité auditive ou visuelle seules (i.e., AV > A + V, cf. Gagné, et al., 2002; Schwartz, et al., 2004, pour des propos similaires).

1.3.1.3. Redondance ou complémentarité de l'information visuelle et auditive ?

Ainsi, le fait voir le visage de son interlocuteur permet d'augmenter l'intelligibilité du signal de parole lorsque celui-ci est détérioré. Or en l'absence de bruit, nous avons vu que l'information auditive est plus spécifiée que l'information visuelle. En effet, le signal acoustique seul suffit pour comprendre le langage oral, alors qu'il a été évalué que l'information visuelle ne permet d'identifier que 10 à 20 % des mots de la langue (e.g., Schwartz, et al., 2004). Ces deux signaux ayant la même origine (i.e., les gestes articulatoires du langage oral), ce constat suggère que les indices traités par la vision sont strictement redondants par rapport à ceux véhiculés auditivement. Or, l'information visuelle est également *complémentaire* du signal acoustique de parole et ce spécialement lorsque le signal acoustique est bruité (e.g., Summerfield, 1987). Plus précisément : « Audiovisual speech complementarity means that one modality is more informative on those dimensions on

[9] « L'apport de l'information visuelle est plus important plus le RSB est faible »

which the other is less informative[10] », (Massaro & Jesse, 2007, p. 20). Cela signifie que l'information issue du signal acoustique de parole est complémentaire à celle issue du signal visuel en termes de *saillance* perceptive. Ainsi, en situation bruitée, ce qui est rapidement masqué auditivement va être facilement perceptible visuellement. Par exemple, l'information relative au voisement (e.g., /p/ vs. /b/) est plus facilement distinguée dans le signal acoustique que visuel, ce dernier résultant de la mise en action d'un organe non visible en situation de communication face à face. Réciproquement, l'information relative à la place d'articulation est très rapidement masquée auditivement par du bruit, alors qu'elle est très saillante visuellement (e.g., Summerfield, 1987). L'étude effectuée par Benoît et al. (1994) (cf. section 1.2.3) apporte également des preuves similaires pour la perception des phonèmes vocaliques. En effet, ces auteurs ont montré qu'en modalité auditive seule, à -12 dB, l'intelligibilité des voyelles (notée I) était ordonnée comme suit : $I_a > I_i > I_y$, signifiant qu'auditivement, un /a/ était plus facilement identifié qu'un /i/ et un /i/ plus facilement identifié qu'un /y/. En modalité audiovisuelle à -24 dB[11], les auteurs ont trouvé que les performances des participants s'ordonnaient plutôt de la manière suivante : $I_y > I_a > I_i$. Ces données viennent donc renforcer cette idée de complémentarité des informations visuelles et auditives puisque qu'auditivement le /i/ semble plus saillant que le /y/ alors que le rapport inverse est observé en modalité audiovisuelle.

En conclusion, l'examen de ces différents travaux montre que la visibilité du visage de notre interlocuteur influence la perception de la parole, lorsque l'information auditive est bruitée (e.g., Sumby & Pollack, 1954). Nous avons mis en évidence que ce bénéfice semble être maximal pour des niveaux de détérioration du signal acoustique intermédiaires (e.g., Ross, et al., 2007). Nous avons également montré que ce gain d'intelligibilité en modalité audiovisuelle semble être supérieur à la somme des performances obtenues en modalité auditive et visuelles seules (i.e., AV > A + V, e.g., Ross, et al., 2007; Schwartz, et al., 2004). Enfin, les différentes études que nous avons évoquées indiquent que le bénéfice lié à la présence de l'information visuelle est observé car lorsque le signal acoustique est détérioré par du bruit, le geste articulatoire de parole n'est pas redondant mais complémentaire par rapport à l'information auditive (e.g., Summerfield, 1987). L'objectif de la prochaine section

[10] « La complémentarité de la parole audiovisuelle signifie qu'une modalité est plus informative sur des dimensions pour lesquelles l'autre modalité est moins informative »

[11] Cette situation correspondait en réalité à une condition de présentation des stimuli en modalité audiovisuelle, avec un RSB de -24 dB. Les performances en modalité auditive seule étant situées autour de 0 %, les auteurs en ont conclu que cette situation équivalait à une situation visuelle seule.

consiste à examiner l'apport de la gestualité oro-faciale à la perception du langage oral en l'absence de détérioration physique du signal acoustique.

1.3.2. Perception audiovisuelle de la parole en l'absence de détérioration du signal acoustique

1.3.2.1. Lorsque les informations auditives et visuelles sont congruentes

Par rapport à une situation où seule l'information auditive est disponible, le fait de voir le visage de son interlocuteur aide également à mieux percevoir le signal de parole (voir Irwin, 2008, pour une revue de littérature), lorsque le discours de notre interlocuteur possède un contenu sémantique complexe (e.g., Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987), est effectué dans une langue étrangère (e.g., Arnold & Hill, 2001; Burfin et al., 2011; Davis & Kim, 2001; Kim & Davis, 2003; Navarra & Soto-Faraco, 2007; Reisberg, et al., 1987; Werker, 2007) ou encore réalisé dans notre langue maternelle mais avec un accent régional différent (e.g., Arnold & Hill, 2001; Reisberg, et al., 1987). Par exemple, les travaux d'Arnold et Hill (2001) ont mesuré que la compréhension d'une histoire présentée en modalité auditive seule ou audiovisuelle chez des participants adultes de langue maternelle anglaise (vivant en Angleterre) et apprenant le français. Les performances observées montrent de meilleurs scores de compréhension d'un discours sémantiquement et syntaxiquement complexe, en modalité audiovisuelle qu'auditive seule, bien qu'il soit prononcé en anglais par un locuteur issu de la même région que les interlocuteurs. Un bénéfice lié à la présence de l'information visuelle a également été observé pour la compréhension d'un texte ne présentant pas de contenu sémantique ou syntaxique particulièrement complexe est présenté en français (quel que soit le niveau de maîtrise de cette langue seconde) ou par un locuteur anglais ayant un accent régional différent de celui des participants (accent écossais). Ainsi, ce travail indique que le fait de voir le visage de son interlocuteur améliore la compréhension de *phrases* entières, même lorsque l'information auditive n'est pas physiquement détériorée par une autre source sonore (bruit).

D'autres travaux ont également étudié ce phénomène pour la perception de traits phonétiques (e.g., Burfin, et al., 2011; Navarra & Soto-Faraco, 2007). Navarra et Soto-Faraco (2007) ont étudié la capacité de discrimination entre deux phonèmes vocaliques existant en espagnol catalan (i.e., /e/ vs. /ɛ/) mais pas en espagnol castillan. Les auteurs ont montré qu'en situations unimodales (i.e., modalité auditive seule, A et visuelle seule, V) des individus bilingues espagnol-catalan à dominance catalane sont capables de percevoir ce contraste

phonétique, alors que des bilingues espagnol-catalan à dominance castillane ne le sont pas. Cependant, leurs résultats indiquent qu'en situation bimodale (i.e., audiovisuelle, AV) les bilingues à dominance castillane arrivent à discriminer perceptivement ces deux productions, même si celle-ci n'existe pas dans leur langue. Ainsi, conformément aux études précédentes, ce résultat met en évidence de meilleures performances pour percevoir un signal de parole en modalité audiovisuelle qu'auditive seule, alors que le signal acoustique est intact. Remarquons que cette étude met encore en avant le pattern de performances AV > A + V, déjà retrouvé dans les travaux précédemment présentés. La raison majeure pour laquelle ce type de pattern est observé sera discutée dans la section suivante. Notons également que ce travail a pu montrer que les participants catalans étaient capables de percevoir un contraste relativement subtil en modalité visuelle seule. En effet, le /e/ est articulatoirement proche du /ɛ/ (cf. Figure 3, Robert-Ribes et al., 1998). Cette capacité vient confirmer notre précédent propos (cf. section 1.3.1.3). En conséquence, en dépit du fait que certaines informations ne soient pas facilement décelables dans le signal visuel seul de parole (e.g., la nasalisation) celles qui sont visibles permettent à un individu de discriminer des contrastes phonétiques relativement fins. Notons ici que telles compétences d'analyse fine de l'information visuelle ont également été mises en évidence chez le nourrisson (e.g., Weikum et al., 2007), (cf. section 5.1.1.1 du Chapitre 5). Enfin, une très récente étude (Burfin, et al., 2011) a montré des résultats similaires aux travaux de Navarra et Soto-Faraco (2007) pour la perception des consones. Pour cela, ces auteurs ont utilisé le contraste phonétique /θ/ vs. /f/ existant en espagnol castillan mais pas en français. Leurs résultats indiquent que des monolingues français avaient de meilleures performances pour discriminer /θe/ de /fe/ en modalité audiovisuelle qu'en modalité auditive seule. Ces résultats, compatibles avec ceux présentés précédemment, indiquent que la présence d'information visuelle améliore la perception de traits phonétiques n'appartenant pas à notre langue maternelle.

En résumé, l'examen de ces études révèle un bénéfice lié à la présence de l'information visuelle lorsque l'information auditive est *intacte*, pour la perception d'un signal de parole présenté dans une langue étrangère (e.g., Arnold & Hill, 2001; Navarra & Soto-Faraco, 2007; Reisberg, et al., 1987), ou ayant un contenu sémantique complexe (e.g., Arnold & Hill, 2001; Reisberg, et al., 1987). Cependant, ces deux situations ne sont pas moins des conditions *difficiles* de perception de la parole. L'objectif de la prochaine section consiste à étudier s'il existe une influence de cette gestualité oro-faciale en présence d'une information auditive intacte, pour une situation où le langage oral est relativement « facile » à comprendre.

1.3.2.2. Lorsque les informations auditives et visuelles sont incongruentes

> « Reality is merely an illusion, albeit a very persistent one[1] »
> (Albert Einstein).

Les illusions perceptives constituent un outil de recherche intéressant puisqu'elles mettent en exergue les différents mécanismes mis en jeu par notre système cognitif pour décoder son environnement. Le domaine de la perception audiovisuelle de la parole n'échappe pas à cette règle. L'illusion perceptive la plus célèbre dans ce domaine de recherche est l'effet McGurk[12]. Celui-ci a été la première fois mise en évidence par McGurk et MacDonald (McGurk & MacDonald, 1976). Ces auteurs ont montré que la présentation d'un signal acoustique /ba/ doublé de la vidéo du visage d'un locuteur articulant un /ga/ était généralement perçue /da/ ou /ða/. Ce phénomène, largement répliqué dans la littérature (voir Colin & Radeau, 2003, pour une revue) constitue une illusion perceptive puisque le résultat observé ne correspond ni au signal visuel ni au signal acoustique, mais à un percept issu de la *fusion* de ces deux sources d'informations. Ainsi, l'effet McGurk montre que l'information visuelle est traitée par notre système visuel, même en l'absence de détérioration du signal acoustique ou de difficulté de perception du langage oral. L'apport primordial de ce phénomène est qu'il met en évidence que les informations visuelles et auditives sont *intégrées* lors de la perception de la parole. Depuis cette découverte, déterminer à quel moment, dans le décours temporel du décodage du langage oral, s'effectue cette intégration est toujours au cœur d'un vaste débat dans la littérature (e.g., Galantucci & Fowler, 2006; Massaro & Chen, 2008). Ainsi, alors que certaines théories supposent que cette intégration s'effectuerait dès les premières phases de traitement (e.g., Galantucci & Fowler, 2006), d'autres postulent qu'elle aurait lieu plus tardivement, chaque information (i.e., auditive et visuelle) étant décodée séparément avant d'être fusionnée l'une à l'autre (e.g., Massaro & Chen, 2008). Dans ce cas, ce prétraitement impliquerait que deux codes séparés (i.e., spécifiques à chaque modalité) soient utilisés pour traiter chaque signal. De ce fait, le décours temporel de cette intégration pose également la question de la *nature* des unités de traitement impliquées lors du décodage du langage oral. Dans la section 1.4 de ce chapitre, nous verrons que ces problématiques divisent l'opinion et opposent les différents modèles décrivant la perception de la parole en modalité audiovisuelle.

Un des avantages de l'illusion McGurk est qu'elle permet de mesurer l'impact de l'information visuelle en modalité audiovisuelle indépendamment de l'information auditive.

[12] Voir aussi, l' « effet de ventriloquie » ou « Ventriloquism effect » (e.g., Vroomen & de Gelder, 2004).

Autrement dit, elle permet d'explorer les mécanismes mis en jeux dans le processus d'intégration, ce dernier pouvant être responsable du pattern AV > A + V observé dans les études évoquées ci-dessus (e.g., Navarra & Soto-Faraco, 2007; Schwartz, et al., 2004). De plus, l'effet McGurk est une illusion robuste et expérimentée dans plusieurs contextes et dans plusieurs langues (cf. Colin & Radeau, 2003). Ces deux avantages en ont fait un outil très fréquemment utilisé pour mesurer l'impact de l'information visuelle sur la perception du signal de parole. Cependant, il n'en résulte pas moins d'une modification artificielle du langage oral. Or, cette situation d'incongruence entre le signal visuel et acoustique de parole est rarement observée dans la vie quotidienne (excepté lors de la perception de films doublés). En effet, à la différence des illusions obtenues pour la perception des objets en modalité visuelle, celle-ci n'est jamais expérimentée qu'en situation de laboratoire. Etudier la perception de la parole exclusivement avec ce type de paradigme poserait un véritable problème de validité écologique.

1.3.3. Décours temporel de la disponibilité de l'information auditive et visuelle dans le signal de parole

Dans les parties précédentes, nous avons présenté une série de travaux visant à déterminer dans quelle mesure l'information visuelle permettait de contribuer à la perception de la parole. Certains travaux se sont également intéressés à déterminer à quel moment, dans le décours temporel de la perception du langage oral, le geste articulatoire contribue à ce processus. Ainsi, nous allons présenter certains travaux ayant étudié la disponibilité et le traitement des signaux acoustiques et visuels au cours du temps, lorsque l'information auditive est intacte (e.g., Cathiard, 1994; Cathiard, Lallouache, Mohamadi, & Abry, 1995; Cox, Norrix, & Green, 1999; Jesse & Massaro, 2010; Munhall & Tohkura, 1998; Seitz & Grant, 1999; Smeele, 1994, voir Cathiard, 1994 pour une revue).

Une des études effectuée dans ce domaine a été menée par Paula Smeele (Smeele, 1994). Elle a montré qu'au niveau de la production d'une syllabe CV (e.g., /pa/), le geste articulatoire pour le phonème consonantique (e.g., /p/) précédait généralement le signal acoustique correspondant (cf. Figure 6). Ainsi, pour certaines syllabes, l'information visuelle serait naturellement et temporellement disponible avant l'information auditive.

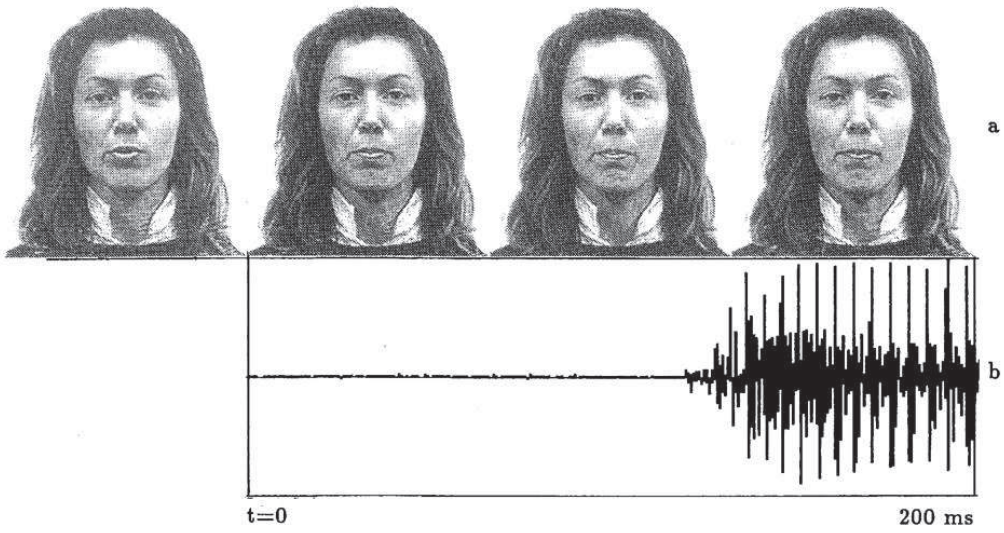


Figure 6. *Illustration du décours temporel des composantes visuelles (a) et acoustiques (b) de parole pour la production de la syllabe /pa/. Le début du signal visuel détermine temps = 0. A t = 200 ms, le stimulus contient de l'information pour la consonne entière et le début de la voyelle /a/. L'intervalle temporel entre chacune des images est de 40 ms. (Extrait de Smeele, 1994).*

Remarquons que pour cet exemple, c'est principalement le mouvement de la fermeture des lèvres pour la consonne occlusive /p/, correspondant à la place d'articulation, qui serait présente dans le signal visuel avant le signal acoustique (100 ms). Afin d'examiner si cette avance dans le signal visuel était exploitée au niveau perceptif, Smeele (1994) a présenté ce stimulus en modalité auditive, visuelle et audiovisuelle à l'aide d'un paradigme de *gating* (Grosjean, 1980). Cette technique consiste à dévoiler une portion de plus en plus large d'un stimulus (i.e., de manière incrémentielle) au fur et à mesure des essais. Elle permet de mesurer le moment (dans le décours temporel du signal de parole) pour lequel un certain type d'information (visuelle, auditive) est disponible et éventuellement traitée par notre système perceptif. A l'aide de cette technique, Smeele (1994) a montré que le phonème /p/ était identifié beaucoup plus tôt dans le décours temporel de la syllabe /pa/ en modalité audiovisuelle et visuelle qu'en modalité auditive seule (cf. Figure 7).

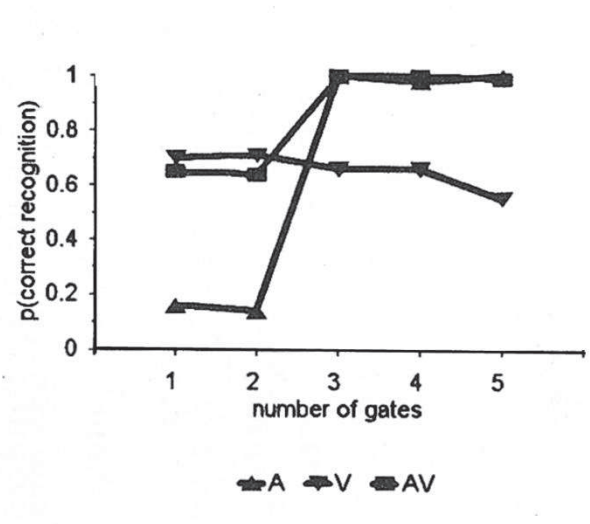


Figure 7. *Pourcentage d'identification correcte du phonème /p/ en fonction de la quantité d'information dévoilée lors de la présentation de la syllabe /pa/ en modalité auditive seule (A), audiovisuelle (AV) et visuelle seule en fonction de la quantité d'information dévoilée. L'intervalle temporel entre chacun des paliers (« gates ») est de 40 ms. (Adapté de Smeele, 1994).*

Ces résultats montrent qu'il peut exister une précédence temporelle de l'information visuelle sur le signal acoustique et que, le cas échéant, l'information qui en est extraite (i.e., la place d'articulation) est exploitée et traitée par notre système cognitif pour percevoir la parole (voir Seitz & Grant, 1999, pour des résultats et un argumentaire similaires avec des mots monosyllabiques). Notons cependant que Smeele (1994) n'a observé ce pattern AV > A que lorsqu'il existait une précédence temporelle de l'information visuelle sur l'information auditive. En effet, pour les syllabes disposant d'une consonne constrictive à l'initiale (e.g., /fa/), aucune avance au niveau de la production du signal visuel n'a été mise en évidence. De facto, un pattern AV ≈ A pour les pourcentages d'identification de ces phonèmes consonantiques a été observé, suggérant que cet avantage est surtout observé pour les syllabes disposant d'une consonne occlusive à l'initiale (e.g., /pa/)[13].

Les travaux effectués par Cathiard (Cathiard, 1994; Cathiard, et al., 1995) ont mis en évidence que cette précédence du signal visuel sur l'information auditive était également exploitée lors de la perception des voyelles. En effet, ce travail a montré que notre système visuel est capable de décoder l'information visuelle (correspondant ici principalement aux mouvements des lèvres d'un locuteur) afin d'identifier un phonème, avant même que toute information auditive relative à ce dernier ne soit disponible dans le signal acoustique. Dans cette étude, le phonème /y/ (condition expérimentale) ou /i/ (condition contrôle) était inséré dans une phrase porteuse (e.g., « Tu dis /y/ ? » vs. « Tu dis /i/ ? »). Cet énoncé a été spécifiquement sélectionné afin d'obtenir une transition vocalique /i → y/ ou / i → i/

[13] Notons que certains travaux (e.g., Troille, Cathiard, & Abry, 2007) ont trouvé le pattern inverse (i.e., A > AV > V) pour la production et le décodage du phonème constrictif /z/.

engendrant une pause (i.e., un silence) dans le signal acoustique (e.g., /tydi.y/, le point indiquant la pause) mais également, pour la condition expérimentale, un changement important de la configuration des articulateurs dans le signal visuel (i.e., principalement des lèvres sur une dimension étirée /i/ → arrondie /y/, cf. Figure 8).

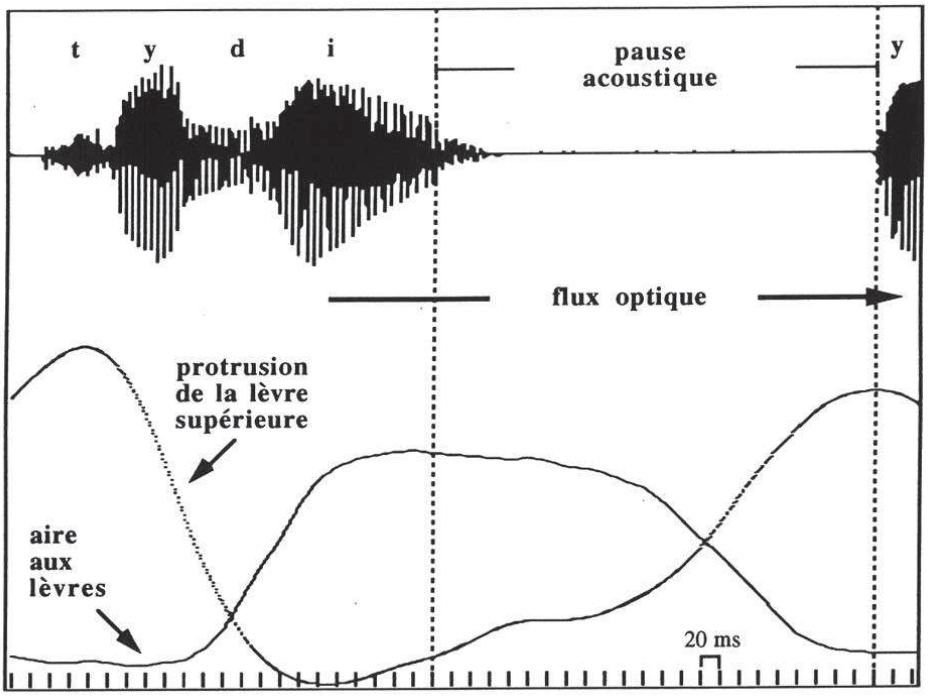


Figure 8. *Signal acoustique correspondant à la phrase « Tu dis /y/ ? » et évolution des paramètres articulatoires de protrusion de la lèvre supérieure et d'aire aux lèvres (i.e., surface entre la lèvre supérieure et inférieure) en fonction du temps. Extrait de Cathiard (1994).*

La phrase porteuse était présentée soit en modalité audiovisuelle, soit en modalité auditive seule. Elle était dévoilée de manière incrémentale aux participants, à l'aide d'un paradigme de *gating*. La tâche était d'identifier la voyelle finale (i.e., /y/ ou /i/). Leurs résultats montrent qu'en modalité audiovisuelle, pour la condition expérimentale, les participants étaient capables d'identifier le /y/ pendant le silence acoustique et donc avant même que toute information auditive relative à ce phonème ne soit disponible dans le signal acoustique (cf. Figure 9).

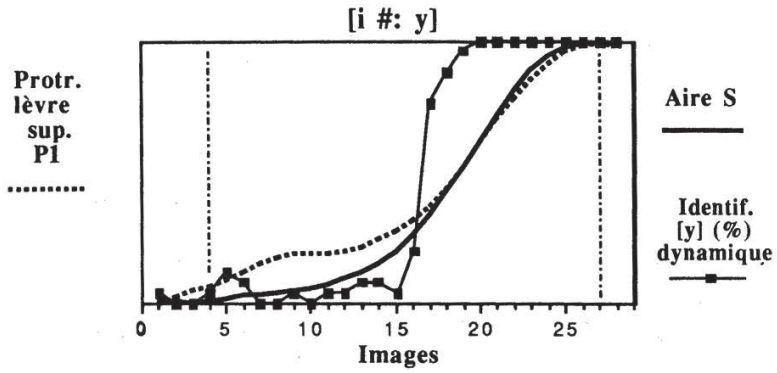


Figure 9. *Pourcentage d'identification du /y/ et évolution des paramètres de protrusion (P1) et d'aire aux lèvres (S) pendant la pause acoustique /i.y/ de la phrase « Tu dis /y/ ? » en fonction de la quantité d'information dévoilée. La verticale point-tiret de gauche indique la fin acoustique du /i/ ; celle de droite le début acoustique du /y/. (Extrait de Cathiard, 1994).*

Par contre, en modalité auditive, les participants n'étaient pas capables d'identifier le phonème qu'après son début acoustique, environ 100 ms plus tard qu'en modalité audiovisuelle. Ce pattern AV > A n'a pas été retrouvé pour la condition contrôle. En conséquence, ces résultats indiquent qu'il existe une précédence du signal visuel sur le signal acoustique pour la production de la transition vocalique /i → y/. De plus, ces données montrent que l'information visuelle relative aux voyelles (i.e., ici, l'arrondissement des lèvres) serait traitée par notre système perceptif avant que le signal acoustique ne soit disponible. Ce traitement visuel du geste articulatoire aurait pour but d'*anticiper* sa conséquence acoustique.

En accord avec cette idée, signalons également les travaux effectués par Munhall et Tohkura (1998). Dans cette étude, les auteurs ont utilisé l'effet McGurk : le stimulus proposé correspondait à la présentation d'un signal acoustique /æbæ/ doublé d'un signal visuel de l'articulation d'un /ægæ/. Le percept illusoire attendu, correspondant à l'intégration de l'information visuelle et auditive était /ædæ/. La tâche demandée aux participants était d'identifier le phonème consonantique situé entre les deux voyelles. Ainsi, le pourcentage de réponse /d/ correspond à la prise en compte de l'information visuelle en présence d'une information auditive parfaitement audible. A l'instar des études précédentes, un paradigme de *gating* en modalité audiovisuelle et visuelle (Expérience 1) ou auditive seule (Expérience 2) a été utilisé. Dans une première condition (Expérience 1), seule l'information visuelle était

dévoilée par pallier, alors que l'information auditive était entièrement présentée (i.e., condition « *gating* visuel »). Dans la seconde condition (Expérience 2), c'est le signal acoustique qui était dévoilée de manière incrémentielle, alors que le signal visuel de parole était conservé dans sa totalité (i.e., condition « *gating* acoustique »). Les résultats concernant le pourcentage d'identification du phonème /d/ sont présentés dans la Figure 10.

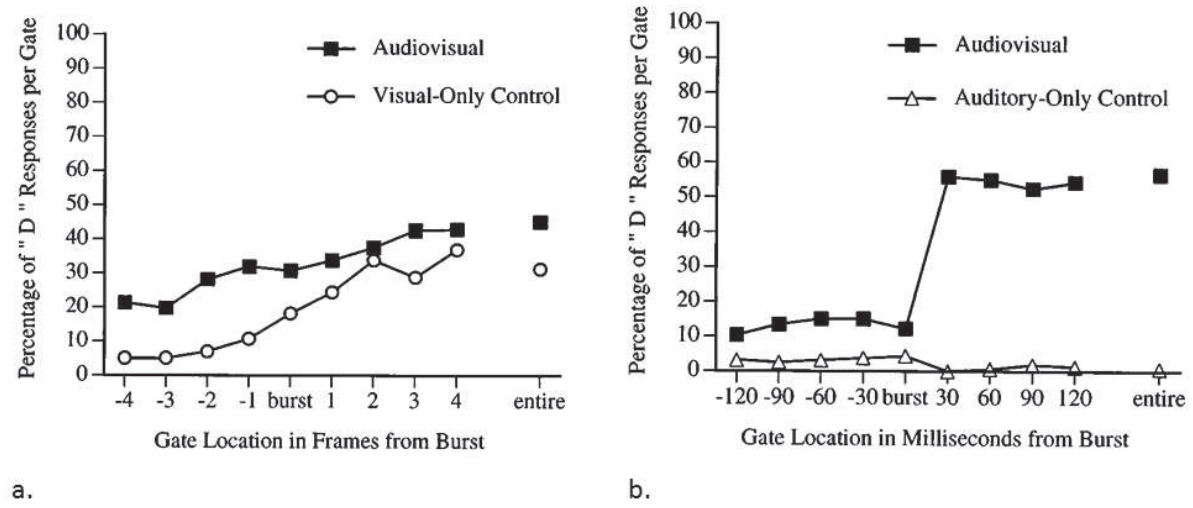


Figure 10. *Pourcentage d'identification du phonème /d/ en fonction de la quantité d'information dévoilée avant et après la plosion acoustique pour les conditions de gating visuel (a) et de gating auditif (b). Ces données ont été récoltées en modalité Audiovisuelle (carrés noirs) Visuelle seule (ronds blancs) et Auditive seule (triangles blancs). (Extrait de Munhall & Tohkura, 1998).*

L'analyse de leurs données montre premièrement que pour la condition *gating* visuel (Figure 10a) en modalité audiovisuelle, le pourcentage de réponse /d/ augmentait de manière *linéaire* en fonction du dévoilement progressif du signal visuel. Cependant, pour la condition *gating* acoustique (Figure 10b) les résultats ont révélé que cette augmentation s'effectuait de manière non linéaire en fonction du dévoilement progressif du signal acoustique ; la contribution de l'information auditive était surtout présente dans la portion CV (i.e., après la plosion acoustique de la consonne) plutôt que VC (i.e., avant la plosion acoustique de la consonne). Ces données mettent en évidence que les informations visuelles et auditives relatives à un phonème ne sont rendues disponibles à un moment et à un rythme différent dans le décours temporel de la parole audiovisuelle. Le signal visuel serait délivré de manière continue lors de la production du langage oral alors que l'information auditive serait rendue disponible dans une fenêtre temporelle beaucoup plus courte et donc de manière beaucoup plus discrète et condensée. Ces résultats, associés au fait que certains traits phonétiques sont présents et traités plus tôt dans le signal visuel que dans le signal acoustique du langage oral (e.g., la place d'articulation, Smeele, 1994), ont amené les auteurs à postuler que l'information visuelle pourrait avoir un rôle d' « amorçage » dans le processus

de reconnaissance de la parole. En d'autres termes, le fait que l' information visuelle soit délivrée à un rythme différent (i.e., plus « continu ») de l'information auditive dans le signal de parole et que certaines d'entre elles soient disponibles et exploitées par notre système perceptif avant l'arrivée du signal acoustique correspondant (Cathiard, 1994; Smeele, 1994) suggère que le geste articulatoire est décodé afin de prédire l'arrivée de l'information auditive.

En accord avec cette hypothèse, Van Wassenhove et collègues (Van Wassenhove, Grant, & Poeppel, 2005) ont observé une réponse cérébrale (i.e., potentiels évoqués) du cortex auditif associatif plus précoce en réponse à la présentation de syllabes CV (i.e., /pa/, /ta/ et /ka/) en modalité audiovisuelle qu'en modalité auditive seule. Notons que leurs résultats ont montré que cette « facilitation » était plus importante pour la syllabe la plus saillante sur le plan articulatoire, c'est-à-dire celle dont le phonème consonantique est articulé à l'avant du conduit vocal (i.e., /pa/). Afin d'expliquer ce phénomène, les auteurs ont proposé que « the natural dynamics of audiovisual speech (e.g., precedence of visual speech inputs) [...] allow the speech processing system to build an on-line prediction of auditory signals[14]" (p. 1185). En d'autres termes, ces auteurs font une hypothèse compatible avec l'idée que l'information visuelle aurait un rôle d' « amorçage » ou de pré-activation de certaines unités impliquées dans le décodage du langage oral. En effet, ils supposent que la quantité d'informations extraites du signal visuel alors même que l'information auditive n'est pas encore disponible permettrait d'initier le processus de traitement de la parole en pré-activant des représentations. Toujours selon ces auteurs, la mobilisation de ces représentations par le signal visuel permettrait d'effectuer des prédictions plus ou moins précises en fonction de sa saillance perceptive, prédictions avec lesquelles l'information auditive serait évaluée, comparée. Ainsi, le décodage de certaines informations présentes très tôt dans le signal visuel (e.g., la place d'articulation) permettrait de prédire/d'anticiper sa conséquence acoustique.

Remarquons néanmoins que toutes les études citées ci-dessus se sont servies d'un très petit nombre de stimuli (i.e., N = 1 à 11). Cela vient du fait que les procédures utilisées ici nécessitent que chaque item soit répété à un même participant un nombre important de fois. Cela a pour conséquence un allongement considérable du temps de passation de l'étude pour chaque participant (par rapport à une situation où chaque item ne lui serait présenté qu'une seule fois). Cependant, utiliser un petit nombre de stimuli engendre que les

[14] « la dynamique naturelle de la parole audiovisuelle (e.g., la précédence de l'entrée visuelle de parole) permet au système de traitement de la parole d'effectuer des prédictions en temps réel du signal auditif »

conclusions tirées de ces études sont difficilement généralisables à l'ensemble du signal de parole. A notre connaissance, le seul travail ayant étudié cette question en utilisant un grand nombre d'items est celui de Jesse et Massaro (Jesse & Massaro, 2010). Pour cela, les auteurs ont utilisé des stimuli synthétiques (parole de synthèse) correspondant à des mots de type CVC. Cela leur a permis de tester de manière exhaustive l'ensemble des consonnes présentes à l'initiale des mots pour la langue anglaise et cela en contrôlant de manière très précise la distribution de l'information visuelle et auditive au cours du temps. Les stimuli étaient présentés soit en modalité auditive, soit en modalité visuelle, soit en modalité audiovisuelle, à l'aide d'un paradigme de *gating*. Les participants devaient identifier le mot correspondant le mieux à la portion initiale de l'item qu'ils avaient perçue. Les auteurs ont ensuite calculé le pourcentage d'information transmise (Shannon, 1948, cité dans Jesse & Massaro, 2010), pour chaque trait phonétique correctement identifié de la consonne initiale. L'examen de leurs données montre que conformément aux études présentées précédemment, les caractéristiques articulatoires les mieux transmises en modalité visuelle seule sont l'arrondissement labial et la place d'articulation (cf. Cathiard, 1994; Smeele, 1994). Ensuite, leurs données montrent que dès le premier pallier, la place d'articulation, l'arrondissement labial, la constriction ainsi que la durée de la consonne, étaient mieux transmis en modalité audiovisuelle qu'auditive seule. Néanmoins, ce bénéfice audiovisuel n'a pas été observé pour le voisement et la nasalisation. Globalement, ces résultats suggèrent donc que le fait de voir le visage son interlocuteur aide à anticiper les conséquences acoustiques des traits articulatoires d'arrondissement-étirement et la place d'articulation en l'absence d'information auditive, mais aussi la constriction et la durée en présence du signal acoustique.

Jusqu'à présent, l'ensemble des travaux présentés ici a employé un paradigme de *gating*. Bien que ce paradigme permette d'évaluer le traitement de l'information à différent point dans un stimulus de parole, cette technique ne permet pas d'analyser directement à quel point dans le temps notre système perceptif est capable d'extraire telle ou telle information. En effet ce paradigme ne permet pas de mesurer la perception d'un individu en temps réel, sur le moment, mais seulement en temps différé, après la présentation du stimulus, généralement avec une tâche d'identification. Ainsi, il est probable que cette technique n'évalue pas seulement les mécanismes automatiques impliqués dans le décodage du signal de parole. En d'autres termes, il est possible que les performances observées avec ce type de tâche soient également « contaminées » par le développement de stratégies conscientes, visant à deviner le stimulus sur la base d'une information partielle (Cutler,

1995). Ainsi, les conclusions tirées des études utilisant ce type de paradigme doivent être considérées avec parcimonie. Le seul travail ayant à notre connaissance étudié le rôle anticipatoire de l'information visuelle avec un autre paradigme et en utilisant des données récoltées en temps réel est celle de Cox et collègues (Cox, et al., 1999). Dans ce travail, les auteurs ont proposé une tâche de détection de phonèmes consonantiques, dans des mots présentés en modalité auditive ou audiovisuelle. Ces mots étaient eux-mêmes insérés dans des phrases porteuses. Leurs données ont montré que les participants étaient plus rapides pour détecter un phonème lorsque le visage du locuteur était visible plutôt que lorsqu'il ne l'était pas. Cela indique donc que la présence de l'information visuelle *accélère* le processus de détection de phonèmes, en présence d'une information auditive congruente et non détériorée. Ce résultat pourrait éventuellement être en accord avec l'hypothèse développée ci-dessus, à savoir que le traitement du geste de parole visible accélère la détection d'un phonème parce que cela permettrait de pré-activer certaines unités (e.g., phonétiques) avant que l'information auditive ne soit traitée. Notons cependant que les données issues de cette étude ne permettent pas de déterminer si c'est la précédence temporelle de l'information visuelle sur l'information auditive ou s'il s'agit simplement d'un bénéfice lié à la présence d'une information supplémentaire (i.e., l'information visuelle) qui a permis la détection plus rapide du phonème en modalité audiovisuelle qu'auditive seule.

En conclusion, plusieurs résultats indiquent que lorsque le signal visuel de parole précède l'information auditive, cette avance temporelle serait exploitée par notre système perceptif afin d'extraire des indices (e.g., la place d'articulation, Smeele, 1994 ; l'étirement-arrondissement, Cathiard, 1994) permettant d'anticiper leur(s) conséquence(s) acoustique(s). Un mécanisme susceptible de rendre compte de ce phénomène serait un traitement du geste articulatoire par notre système visuel afin de pré-activer certaines unités (e.g., phonétiques, Munhall & Tohkura, 1998) permettant de prédire l'arrivée d'un signal acoustique de parole congruent (Van Wassenhove, et al., 2005). Ainsi, la mise en action de ce mécanisme permettrait de reconnaître plus rapidement un phonème (Cathiard, 1994; Cox, et al., 1999; Jesse & Massaro, 2010; Munhall & Tohkura, 1998; Smeele, 1994), une syllabe (Van Wassenhove, et al., 2005) ou encore un mot (Seitz & Grant, 1999) lorsque l'information visuelle est disponible dans la situation de communication (i.e., en modalité audiovisuelle), par rapport à une situation où seul le signal acoustique de parole est disponible (i.e., en modalité auditive). Or, une des hypothèses sous-tendant le travail présenté dans ce manuscrit concerne le type de représentation que l'information visuelle permettrait d'activer afin de prédire l'arrivée de l'information auditive. En effet, les études citées ci-dessus ont

exploré le rôle anticipatoire de l'information visuelle ; la majorité d'entre elles postule que des unités phonétiques ou phonémiques (i.e., pré-lexicales) étaient pré-activées par l'information visuelle alors que les autres travaux restaient vagues quant au type de représentations impliquées dans ce mécanisme. L'idée générale ayant motivé le travail présenté dans ce manuscrit, est de postuler que l'information visuelle ne permettrait pas seulement de pré-activer des représentations pré-lexicales mais également d' « amorcer » les représentations lexicales (i.e., relatives aux mots). Ainsi, décoder l'information visuelle permettrait non seulement d'anticiper des informations d'ordre phonétique mais également de pré-activer un certain nombre de représentations de mots compatibles avec le signal visuel d'entrée. Cette hypothèse sera plus amplement développée dans la section 6.2.2.1 du Chapitre 6. L'objectif de la partie suivante consiste à brièvement présenter les principaux modèles décrivant la perception de la parole en modalité audiovisuelle.

1.4. MODELES DE LA PERCEPTION AUDIOVISUELLE DE LA PAROLE

Le but de cette section consiste à présenter de manière non exhaustive un certain nombre de modèles considérant la parole comme un évènement bimodal (i.e., audiovisuel). Nous tenons à faire remarquer au lecteur que ces modèles ne sont proposés qu'à titre indicatif, puisque l'objet de ce travail de thèse ne consiste pas à départager ni même à prendre parti pour l'un d'entre eux (voir Basirat, 2010; Schwartz, Robert-Ribes, & Escudier, 1998; Schwartz, Sato, & Fadiga, 2008, pour des revues complètes à ce sujet). Notons également que les modèles présentés ici correspondent uniquement à ceux qui postulent explicitement une fusion entre les informations auditives et visuelles et qui sont toujours vivement débattus dans la littérature (e.g., Galantucci & Fowler, 2006; Massaro & Chen, 2008). Certains d'entre eux postulent une intégration « tardive » sur l'hypothèse qu'une catégorisation phonétique précèderait la fusion des informations auditives et visuelles. D'autres supposent une intégration précoce qui s'effectuerait dès les premières phases de traitement du langage oral et s'opèrerait sur des représentations communes aux deux modalités. Nous avons donc choisi d'utiliser préférentiellement cette dichotomie (i.e., intégration tardive vs. précoce) pour présenter les différents modèles de perception audiovisuelle de la parole. Notons qu'il existe des classifications plus complexes (e.g., Schwartz, et al., 1998; Summerfield, 1987) prenant en compte d'autres paramètres (e.g., présence/absence de codage commun aux deux modalités). Ces caractéristiques n'étant pas directement liées à notre question de recherche, nous nous sommes contentés de les citer sans les prendre en considération dans la typologie des différents modèles décrits ici. Nous

achèverons cette section par la présentation d'un modèle considérant la parole audiovisuelle dans une perspective perceptivo-motrice.

1.4.1. Modèles de perception de la parole à intégration « tardive »

Plusieurs modèles supposent que les informations visuelle et auditive seraient traitées séparément avant d'être intégrées l'une à l'autre « Lexical Access from Spectra and Face Parameters[15] », LASFP, (Klatt, 1979, cité par Klatt, 1989), « Fuzzy Logical Model of Perception », FLMP, (Massaro, 1998), « Vision Place Audition Manner[16] », VPAM, (McGurk & MacDonald, 1976). Nous avons choisi de présenter le fonctionnement du FLMP, car il semble être celui qui résiste le plus à ses concurrents. Dans ce modèle, lors d'une première étape d'évaluation (cf. Figure 11), chacune des sources d'information (e.g., le flux visuel noté V_j le signal acoustique noté A_i) est évaluée séparément. Chacune de leurs propriétés est comparée à plusieurs prototypes stockés en mémoire (e.g., pour un /b/ : la fermeture des lèvres pour la modalité visuelle, la plosion acoustique pour la modalité auditive).

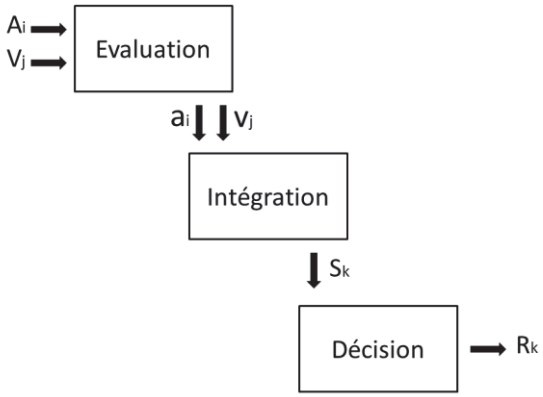


Figure 11. *Représentation schématique des trois processus principaux impliqués dans la reconnaissance de la parole pour le FLMP. (Adapté de Massaro, 1998).*

Une valeur (« fuzzy truth value[17] ») est assignée à chaque propriété (ou « trait ») issue de chaque source d'information (notées a_i pour celle associée à l'entrée auditive, v_j pour celle associée à l'entrée auditive), en fonction de l'adéquation d'un signal avec le « prototype » mémorisé. Cette valeur peut être comprise entre 0 et 1, une valeur de 0.5

[15] Littéralement : « Accès lexical à partir des paramètres du spectre et du visage »
[16] Littéralement : « Vision place audition mode ». Cette théorie postule que certaines informations phonétiques (e.g., la place d'articulation) seraient majoritairement transmises par le signal visuel alors que d'autres (e.g., le mode d'articulation) serait principalement transmis par le signal acoustique de parole. Voir Summerfield, 1987, pour une remise en question de ce modèle.
[17] Littéralement : « valeur de vérité floue »

correspondant à une ambiguïté complète, celle de 0.7 à une adéquation plus vraie que fausse, etc. Ce n'est que dans une deuxième étape (i.e., étape d'intégration) que les propriétés de chacune des modalités sont intégrées. Le poids accordé à l'une ou l'autre modalité dans le résultat de l'intégration est modélisé avec cette valeur. Après cette phase, un degré d'adéquation général (noté S_k) de l'entrée sensorielle à chacun des prototypes est calculé. Une étape de décision permet enfin au système de sélectionner le prototype le plus cohérent avec les informations visuelles et auditives codées.

En conséquence, le FLMP postule clairement deux processus de décodage séparés de l'entrée auditive et visuelle pendant lesquels des caractéristiques phonétiques sont extraites séparément de chaque signal, *avant* que celles-ci ne soient fusionnées. En d'autres termes, l'information issue d'une modalité est évaluée et catégorisée séparément avant d'être intégrée aux informations véhiculées par l'autre entrée sensorielle. C'est en ce sens que nous classons ce modèle parmi ceux postulant une intégration tardive (cf. Schwartz et al., 1998, pour une proposition similaire). Ainsi, avant l'intégration, les informations visuelles et auditives sont chacune traitées avec un code spécifique à la modalité d'entrée. Les valeurs associées ensuite à chacune des entrées perceptives correspondent au code commun postulé par ce modèle permettant la réalisation d'une fusion entre ces deux informations. Notons que ce modèle postule également que l'intégration entre les informations visuelles et auditives s'effectuerait selon une loi fixe, c'est-à-dire indépendamment de facteurs attentionnels, contextuels ou inter-individuels (i.e., des spécificités liées aux individus).

1.4.2. Modèles de perception de la parole à intégration « précoce »

La Théorie Motrice de la Perception de la Parole ou MTSP (« Motor Theory of Speech Perception », Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Liberman & Whalen, 2000) suppose quant à elle que les informations auditive et visuelle seraient intégrées précocement dans le processus de perception de la parole. Ainsi, contrairement au FLMP, elle postule que nous effectuons un codage *amodal* des entrées visuelles et auditives et que ces dernières sont intégrées précocement. Comme son nom l'indique, la nature de ce codage ne serait pas liée à la modalité perceptive mais serait associée aux caractéristiques articulatoires du langage. En d'autres termes, la MTSP postule que les informations issues de la vision (i.e., les mouvements de la face) ainsi que celles véhiculées par le canal auditif (i.e., le signal acoustique), seraient codées en termes de gestes articulatoires par notre système perceptif. Remarquons que dans cette version de la MTSP, ce ne sont pas les réels mouvements des articulateurs mais plutôt les gestes

intentionnels (« intended gesture ») qui sont codés. Notons également que cette théorie postule l'existence d'un système de perception inné et spécifique au langage.

Ce modèle a été tout d'abord élaboré pour résoudre une problématique bien connue dans la perception du langage : celle de l'invariance (cf. section 2.2.1 du Chapitre 2 pour plus d'explication de ce phénomène). En effet, de nombreuses sources de variation peuvent modifier le signal acoustique. Par exemple, le phénomène de coarticulation renvoie au fait qu'un son consonantique n'a pas les même caractéristiques (en fréquence et en amplitude) en fonction du phonème vocalique qui lui succède (cf. Figure 12). Cependant, au niveau perceptif, nous n'avons aucune difficulté à reconnaître le phonème /d/ dans la syllabe /di/ ou dans la syllabe /du/. Ainsi, la théorie motrice de la perception suppose qu'un interlocuteur récupère l'invariance contenue dans les commandes neuro-motrices transmises aux articulateurs (e.g., l'intention d'effectuer l'occlusion pour la réalisation du /d/) pour désambiguïser le signal acoustique.

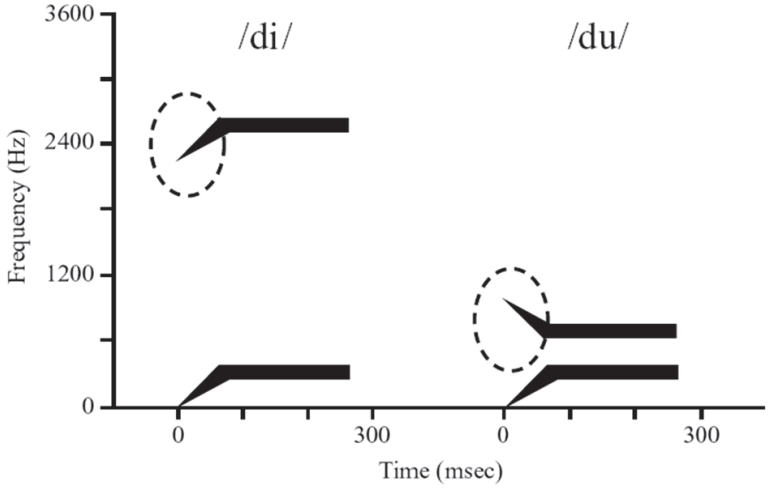


Figure 12. *Formants F_1 et F_2 pour les syllabes synthétiques /di/ et /du/. Les traits pointillés montrent que la transition formantique de F_2 est différente pour /di/ et pour /du/ alors que celle-ci indique la même information sur la place d'articulation alvéodentale). (Extrait de Galantucci & Fowler, 2006).*

Notons que la MTSP est à la base de la théorie formalisée par Fowler en 1986, intitulée la « Théorie Réaliste de la perception Directe de la parole» (« Direct Realist Theory of speech perception », Fowler, 1986, 1996; Galantucci & Fowler, 2006) qui suppose un accès direct du monde par les sens. Cette théorie se distingue de la précédente sur deux points. Elle suppose tout d'abord que ce ne sont pas les gestes moteurs intentionnels mais bien les mouvements articulatoires *véritables* qui sont codés par notre système perceptif. Le second aspect sur lequel Fowler se distingue de ses prédécesseurs est le suivant : elle postule que le module responsable du traitement de la parole ne serait pas spécifique au langage et

ainsi que les mêmes mécanismes seraient à l'œuvre pour la perception dans d'autres modalités sensorielles.

Schwartz et collègues (Schwartz, et al., 2008) font remarquer que bien que la popularité de ces théories motrices ait déclinée dans les années 90, celles-ci ont suscité un regain d'intérêt avec la découverte des « neurones miroirs » (e.g., G. Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). Ces auteurs ont permis de mettre en évidence que des neurones situés dans le cortex pré-moteur du macaque répondaient aussi bien lors de l'*exécution* par le singe d'une action spécifique (e.g., saisir un objet) que lors de l'*observation* par le macaque de l'exécution d'une action similaire effectuée par l'expérimentateur. En d'autres termes ces auteurs ont permis d'apporter des preuves neuro-anatomiques indiquant que certaines zones cérébrales seraient aussi bien recrutées pour la *perception* d'une activité que lors de la *production* de celle-ci. D'autres études ont également mis en évidence l'implication du système moteur chez l'homme, lors de la perception du langage oral (e.g., Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Pulvermüller et al., 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004). Par conséquent, bien que l'existence de « neurones miroirs » chez l'homme soit sujet à débat dans la littérature (e.g., Hickok, 2009; Giacomo Rizzolatti & Craighero, 2004) cette hypothèse est en faveur de l'implication des représentations motrices dans la perception de la parole (Galantucci & Fowler, 2006). Cependant, plusieurs auteurs (e.g., Remez, 1996, 2005; Schwartz, Basirat, Ménard, & Sato, 2010) envisagent à présent la parole comme un objet ni purement moteur ni purement perceptif, permettant de fournir un compromis entre un codage uniquement perceptif ou moteur. La prochaine section consiste donc à présenter une des théories perceptivo-motrices de la parole.

1.4.3. La théorie de la perception pour le contrôle de l'action (PACT)

La théorie de la perception pour le contrôle de l'action (« Perception-for-Action-Control Theory », Schwartz, et al., 2010) considère la parole comme un objet multimodal impliquant la mise en action d'unités perceptivo-motrices. La parole est considérée dans une perspective perception-production, c'est-à-dire à l'interaction de ces deux activités et non séparément l'une de l'autre. Cette théorie est régie par deux concepts centraux : la production du langage oral est influencée par la perception de celle-ci (« Perception shapes action ») et réciproquement, la perception de la parole est contrainte par l'action (« Action shapes perception »).

Les auteurs justifient le besoin de représentations perceptivo-motrices en prenant comme exemple l'arrondissement des lèvres consécutif à la transition vocalique /i/ → /y/.

Ils soulignent que l'exécution articulatoire /i/ → /y/ se réalise progressivement et de manière continue alors que le changement perception du /i/ vers le /y/ va s'effectuer de manière discrète. Ce phénomène de perception catégorielle fait donc référence à une relation non linéaire entre la production et la perception du signal de parole. Cela indique que plusieurs configurations motrices vont pouvoir correspondre à un événement perceptif. En d'autres termes, les changements de position des articulateurs de la parole ne vont pas forcément donner lieu à une perception différente. Par conséquent, celle-ci ne peut être uniquement supportée par des représentations motrices. Cela illustre le premier concept selon lequel la perception influence la production du geste de parole : « Perception shapes action ». Le second principe « Action shapes perception » est supporté par des arguments neuro-anatomiques (i.e., existence de « neurones-miroirs » pour la parole chez l'homme, cf. section ci-dessus) mais aussi par des preuves issues d'études comportementales (e.g., Rosenblum, Miller, & Sanchez, 2007, voir Galantucci et al., 2006, Rosenblum, 2005, 2008, pour des revues). Ces travaux suggèrent l'existence de connections entre les systèmes de perception et de production de la parole, qui permettent au système moteur d'être activé lors de la perception d'un son de parole.

Ce modèle permet, à l'instar des modèles évoqués dans cette section, de rendre compte de la perception de la parole en modalité audiovisuelle et postule l'existence d'un mécanisme de fusion des informations auditives et visuelles. Sa singularité vient du fait qu'en plus de ce processus d'intégration, la PACT fait également l'hypothèse d'un mécanisme de « liage » (cf. Figure 13).

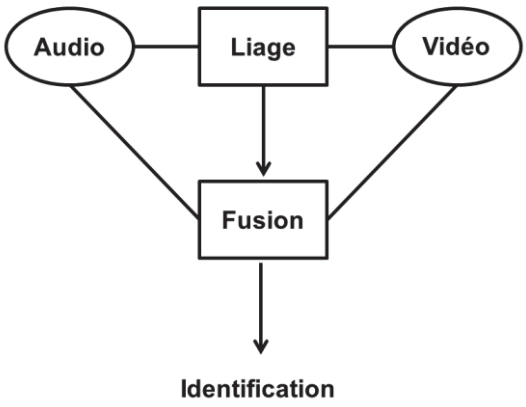


Figure 13. *Mécanisme de liage audiovisuel permettant ou non la fusion entre les informations auditives et visuelles. (Adapté de Basirat, 2010)*

Ce liage interviendrait de manière précoce, préalablement à la fusion des informations auditives et visuelles et conditionnerait directement son existence. Il aurait pour rôle de

regrouper les éléments auditifs et visuels présents dans les signaux de parole afin d'obtenir des percepts audiovisuels cohérents. Ce mécanisme permettrait de rendre du compte du fait de certains résultats observés dans la littérature, comme ceux suggérant que la fusion entre les informations auditives et visuelles n'est pas automatique et dépend notamment de facteurs attentionnels (e.g., Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Tiippana, Andersen, & Sams, 2004) contextuels (e.g., Nahorna, Berthommier, & Schwartz, 2010) et inter-individuels (e.g., Schwartz, 2010). Cette hypothèse de liage est également soutenue par d'autres travaux qui indiquent que la présence d'une information visuelle permettrait de guider l'extraction de l'information auditive pertinente dans le signal acoustique. Autrement dit, le geste articulatoire nous permettrait de segmenter le signal de parole en unités discrètes par ce système de liage (voir section 6.2.2.2 du Chapitre 6 pour une plus ample description de ces travaux).

1.5. BASES NEURALES DE LA PERCEPTION VISUELLE ET AUDIOVISUELLE DE LA PAROLE

L'objectif de cette section consiste à présenter les différentes structures cérébrales qui semblent spécifiquement impliquées dans la perception de la parole visuelle et audiovisuelle chez l'Homme. Cette section n'ayant pas pour ambition d'effectuer un état de l'art à ce sujet, nous invitons le lecteur à consulter des revues complètes décrivant les régions cérébrales impliquées dans le traitement du langage (e.g., Démonet, Thierry, & Cardebat, 2005), de la parole en modalité auditive (e.g., Bernstein, 2005; Hickok & Poeppel, 2007) audiovisuelle et visuelle seule (Möttönen, 2004; Ojanen, 2005).

1.5.1. Régions cérébrales spécifiquement impliquées dans la perception de la parole en modalité visuelle

Chez l'être humain adulte, la première zone cérébrale ayant été mise en relation avec la parole a été découverte par Paul Broca en 1861. Ce médecin a mis en évidence qu'une région corticale située dans l'hémisphère gauche au niveau du gyrus frontal inférieur (désormais appelée « aire de Broca », cf. Figure 14) était impliquée lors de la production de la parole. Quelques années plus tard, Carl Wernicke a permis de montrer que la zone cérébrale située dans la partie postérieure du lobe temporal supérieur de l'hémisphère gauche (également appelée « aire de Wernicke » cf. Figure 14) jouait un rôle dans la compréhension du langage oral. Depuis ces découvertes, plusieurs études se sont intéressées aux zones

spécifiquement impliquées dans le décodage du signal visuel de parole. Ainsi, outre les aires visuelles primaires, il semblerait que plusieurs régions corticales soient impliquées dans cette activité. Un des premiers travaux ayant étudié cette question a été effectué par Calvert et al. (Calvert et al., 1997), dans une étude par Imagerie par Résonnance Magnétique fonctionnelle (IRMf). Dans ce travail, des participants ne présentant aucun déficit auditif ou visuel particulier devaient observer un locuteur articuler silencieusement des nombres. Les résultats montrent une activation (dans l'hémisphère gauche principalement) de la partie postérieure du Gyrus Temporal Supérieur (GTS), où se situe le cortex auditif primaire (situé dans le lobe temporal, cf. Figure 14).

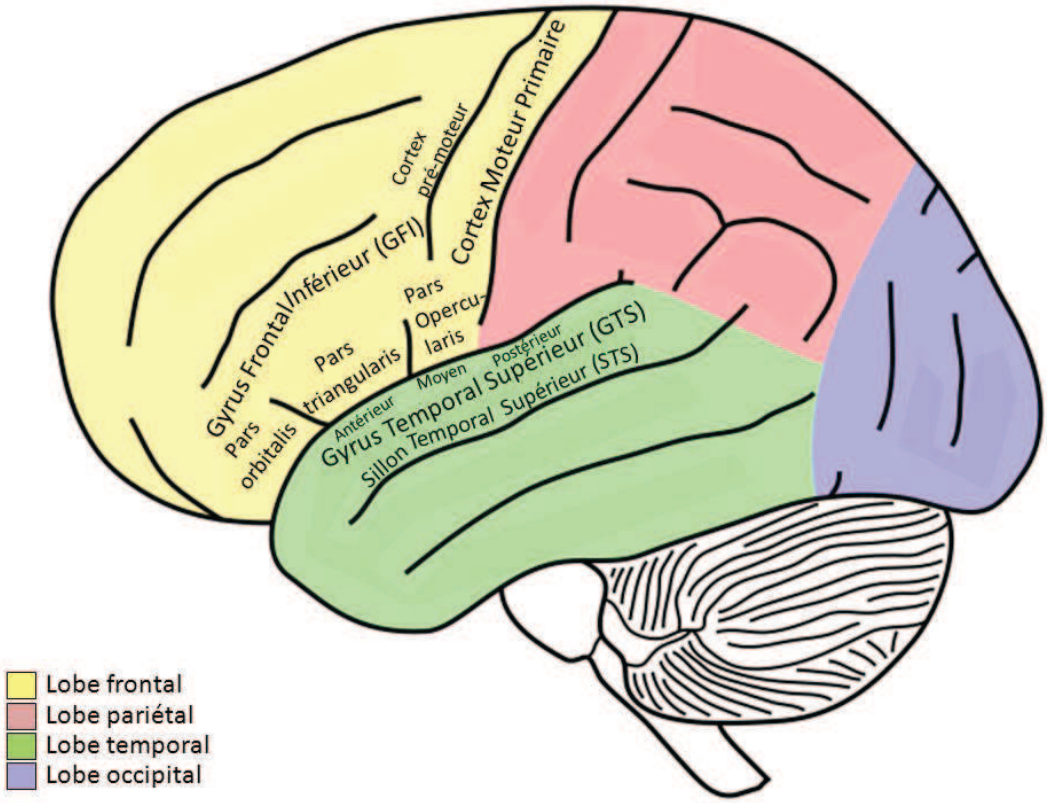


Figure 14. *Représentation schématique des principales régions corticales impliquées dans le décodage du signal visuel et audiovisuel de parole.*

Plusieurs études ont retrouvé ce pattern de résultats, à savoir une activation spécifique des régions auditives dans la partie postérieure du cortex temporal supérieur lors d'une activité de lecture labiale, en l'absence de toute information auditive (e.g., Campbell et al., 2001; MacSweeney et al., 2000). Toujours en situation visuelle seule, il a également été mis en évidence que le Sillon Temporal Supérieur (STS) était spécifiquement activée lors de la présentation de mouvements articulatoires de parole (e.g., articulation silencieuse d'un /u/) mais pas lors de la perception de gestes faciaux similaires n'appartenant pas au langage

oral (i.e., « gurning », e.g., mouvement de protrusion des lèvres sans ouverture de la bouche, cf. Calvert, et al., 1997; Campbell, et al., 2001). En accord avec ces résultats, d'autres travaux ont montré que la perception de points lumineux représentant un individu en train de marcher n'activait pas le STS dans l'hémisphère gauche de la même manière que la perception de points lumineux modélisant le visage d'une personne en train de parler (Santi, Servos, Vatikiotis-Bateson, Kuratate, & Munhall, 2003). Par conséquent, il semblerait que la partie postérieure du STS de l'hémisphère gauche soit impliquée dans le traitement du signal visuel de parole chez des individus normo-entendants. Comme la lecture labiale ne semble pas activer ces mêmes zones chez des individus sourds congénitaux (MacSweeney, et al., 2000), cela suggère qu'une expérience auditive et/ou audiovisuelle de la perception de la parole est nécessaire pour que ces régions soient recrutées lors de la perception de la parole en modalité visuelle seule.

Il a également été observé que des régions corticales impliquées dans la planification et l'exécution de la production de la parole (e.g., l'aire de Broca, le cortex pré-moteur) étaient activées dans des tâches de lecture labiale (e.g., Callan et al., 2003; Campbell, et al., 2001; MacSweeney, et al., 2000; Santi, et al., 2003). Ces données semblent donc indiquer que la perception du geste de parole est médiatisée par l'activité des neurones impliqués dans sa production (e.g., Wilson, et al., 2004). Par conséquent, ces résultats plaident en faveur de l'implication du système moteur dans la perception de la parole (e.g., Galantucci & Fowler, 2006; Schwartz, et al., 2008). En accord avec cette hypothèse, une étude effectuée avec une technique de Stimulation Magnétique Transcrânienne (« Transcranial Magnetic Stimulation », TMS) a même montré que la perception d'un visage articulant silencieusement de la parole modulait le fonctionnement du cortex moteur primaire, spécifiquement dans la zone responsable des mouvements de la bouche dans l'hémisphère gauche (Watkins, Strafella, & Paus, 2003).

1.5.2. Régions cérébrales spécifiquement impliquées dans la perception de la parole en modalité audiovisuelle

Plusieurs travaux ont également étudié les zones corticales semblent spécifiquement impliquées dans le décodage d'un stimulus de parole en modalité audiovisuelle, par rapport à une situation où seule la composante acoustique du langage oral est proposée (i.e., en modalité auditive seule). Ainsi, certaines études semblent indiquer que la partie postérieure du STS (e.g., Calvert, Campbell, & Brammer, 2000; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003) est également impliquée dans la perception du langage oral en modalité

audiovisuelle. Par exemple, les travaux de Calvert et al. (2000) effectués en IRMf ont montré que cette région répondaient plus fortement à des stimuli de parole présentés en modalité audiovisuelle (AV), avec des informations auditives et visuelles congruentes) qu'en modalité auditive seule (A). Ces auteurs ont également trouvé que la quantité d'activation observée en condition AV était supérieure à la somme des activités mesurées en conditions unimodales A et visuelle seule (V). Ce pattern supra-additif AV > A + V a également été retrouvé dans des études comportementales (e.g., Navarra & Soto-Faraco, 2007; Schwartz, et al., 2004) (cf. section 1.3.1.2) et est interprété comme le résultat de l'intégration entre les informations auditives et visuelles. De manière intéressante, lorsque les informations auditives et visuelles étaient incongruentes, un pattern sub-additif a été retrouvé (i.e., AV < A + V). Ces résultats ont amené Calvert et al. (2000) à désigner le STS comme une structure cérébrale jouant un rôle clé dans le processus de fusion de ces deux informations. En accord avec cette idée, des données récoltées à l'aide de la TMS ont permis à l'équipe de Michael Beauchamp de montrer qu'une seule impulsion magnétique dans la zone du STS interférait avec la perception de l'effet McGurk, mais pas avec le traitement de stimuli audiovisuels de parole congruents (Beauchamp, Nath, & Pasalar, 2010). Ce résultat vient conforter l'idée que le STS jouerait un rôle critique dans l'intégration audiovisuelle des signaux de parole (cf. Hocking & Price, 2008, pour une critique vis-à-vis de cette hypothèse). Notons que rôle du STS a également été étudié lors de la perception de la parole en modalité audiovisuelle lorsque l'information auditive est détériorée par du bruit (e.g., Callan, et al., 2003). Dans une étude effectuée en IRMf, Callan et collaborateurs ont par exemple mis en évidence une plus grande activité dans les régions du STS lorsqu'un signal de parole est présenté en modalité audiovisuelle avec du bruit plutôt que sans bruit dans le signal acoustique. Ainsi, ce résultat suggère que cette région serait également recrutée pour traiter l'information visuelle, permettant d'augmenter l'intelligibilité du signal acoustique lorsque ce dernier est bruité.

Remarquons que plusieurs travaux indiquent également que les représentations motrices semblent impliquées dans le décodage d'un signal de parole en modalité audiovisuelle (e.g., Callan, et al., 2003; Calvert, et al., 2000; Ojanen, 2005). En effet, Callan et al. (2003) ont par exemple montré que le cortex pré-moteur et l'aire de Broca étaient plus fortement activés lors de la perception de stimuli audiovisuels que lors de la perception de stimuli auditifs seuls. Ces régions étant impliquées dans la planification et l'exécution de la production de la parole, cela suggère que la perception de l'information visuelle en présence de l'information auditive pourrait être supportée (du moins en partie), par les régions motrices responsables de la production du langage oral.

En conclusion, la présentation de ces études suggèrent que des rôles centraux sont joués à la fois par le système perceptif (e.g., STS) et le système moteur (e.g., le cortex pré-moteur, l'aire de Broca, etc.) dans le décodage de la parole en modalité auditive, visuelle et audiovisuelle. Egalement, il semblerait que ces régions soient fortement impliquées dans le mécanisme d'intégration des informations auditives et visuelles lors de la perception audiovisuelle de la parole (cf. Hickok & Poeppel, 2007, pour une proposition d'architecture anatomique et fonctionnelle de ce phénomène).

1.6. CONCLUSIONS

L'examen des différents travaux présentés de ce chapitre nous a permis de souligner que la perception de la parole n'est pas un évènement purement auditif mais bien un objet multimodal. En effet, en situation de communication face à face, voir le visage de son interlocuteur permet d'augmenter l'intelligibilité d'un signal acoustique détérioré par du bruit (e.g., Sumby & Pollack, 1954). La présence de cette information améliore le décodage de la parole lorsque celle-ci est difficile à comprendre (e.g., Reisberg, et al., 1987). De plus, en l'absence de toute difficulté inhérente au contenu ou à la qualité du signal acoustique, voir un geste articulatoire contraint la perception des sons jusqu'à créer des illusions perceptives (i.e., effet McGurk, McGurk & MacDonald, 1976). La présence de cet effet démontre que les composantes acoustiques et visuelles du signal de parole sont intégrées de manière non consciente par notre système perceptif. Bien que le caractère automatique de ce phénomène soit aujourd'hui questionné (e.g., Alsius, et al., 2005) cette découverte a suscité l'élaboration de plusieurs modèles théoriques et anatomo-fonctionnels décrivant la perception de la parole en modalité audiovisuelle (e.g., Hickok & Poeppel, 2007; Liberman & Whalen, 2000; Massaro, 1998; Schwartz, et al., 2010). Cependant, la mise en action de certains articulateurs n'étant pas visible, la perception de certains contrastes phonétiques est rendue impossible en modalité visuelle seule (e.g., le voisement). Or, le décodage de ces contrastes est essentiel à la différenciation de certaines unités de taille plus importante comme le phonème (i.e., différence entre un /b/ et un /p/) ou le mot (e.g., /bɛ̃/, « bain » vs. /pɛ̃/, « pain »). En conséquence, seul un faible pourcentage de mots peut être consciemment identifié sur la seule base des gestes articulatoires de parole. Ainsi, alors qu'un grand nombre de travaux a cherché à quantifier et à caractériser l'information d'ordre pré-lexical (i.e., faisant référence à des unités plus petites que le mot) pouvant être extraite de ce signal, moins d'études ont examiné la contribution spécifique de l'information visuelle au processus de reconnaissance de mots per se (i.e., unités *lexicales*).

Curieusement, la plupart des travaux décrivant le décodage de la parole au niveau lexical a majoritairement étudié ce phénomène en modalité auditive seule. L'objectif du prochain chapitre (Chapitre 2) consiste donc à passer en revue les différentes études et modèles décrivant ce phénomène, avant d'essayer de répondre à notre propre question de recherche (Chapitre 3 à 6). En effet, l'objectif du travail rapporté dans ce manuscrit est d'appréhender le rôle joué par le signal visuel de parole dans le processus d'accès au lexique.

CHAPITRE 2. LE PROCESSUS DE RECONNAISSANCE DE MOTS : DU SIGNAL AU LEXIQUE

« Why are psycholinguists interested in spoken word recognition? Imagine a typical listening situation. The phone rings, and you find yourself being addressed by an unknown speaker. […] Because each sentence that you hear comes from an unlimited set of potential sentences, it would be impossible to derive what speakers mean by trying to recognize their utterances as wholes. But utterances are made from a limited set of words that, for fluent speakers of a language, will usually already be stored in long-term memory. So speakers'messages must be decoded via recognition of their part."

James M. McQueen, (2007), p. 37.

2.1. INTRODUCTION

Dans le Chapitre 1, nous avons présenté plusieurs théories décrivant les premières phases du décodage de la parole. Ces dernières se sont concentrées sur les étapes précoces de traitement du langage oral, correspondant à une analyse de « bas niveau[18] » (e.g., phonétique, phonémique) de ce signal. Cependant, bien qu'utiles à la perception de la parole, ces unités de bas niveau ne sont pas porteuses de signification lexicale. Or, la fonction première d'une activité linguistique et de la communication en général consiste justement à transmettre du *sens*. Dans le domaine du langage parlé, le *mot* constitue la plus petite unité indépendante (i.e., pouvant être prononcée isolément) capable de véhiculer de l'information d'ordre sémantique (e.g., Altmann, 1997). Ainsi, un grand nombre de recherches, majoritairement effectuées dans le domaine de la psycholinguistique, se sont spécifiquement intéressées aux processus de « haut niveau » (e.g., relatifs aux mots) impliqués dans la perception du langage oral. Dans la suite de ce chapitre, nous allons donc passer en revue différents travaux et modèles décrivant le processus de reconnaissance de mots (i.e., l'accès au lexique) en modalité auditive seule.

L'étude de ce phénomène repose sur plusieurs constats, suppositions et problèmes. Premièrement, nous verrons que le signal acoustique du langage oral dispose de caractéristiques nécessitant, selon différents modèles psycholinguistiques actuels, l'existence de représentations abstraites de mots pour permettre un décodage efficace de la parole. A travers la présentation de différents travaux, nous essaierons de caractériser quels types de représentations sont impliqués lors de l'accès au lexique et à quel moment dans le décours temporel du mot. Ensuite, nous exposerons différentes études montrant l'influence de ces représentations abstraites sur le décodage de l'information linguistique. Nous présenterons alors différents modèles décrivant le processus de reconnaissance de mots en modalité auditive, permettant de rendre plus ou moins compte de l'ensemble de ces données. Nous terminerons ce chapitre en présentant brièvement les structures cérébrales qui semblent spécifiquement impliquées dans le processus de reconnaissance de mots en modalité auditive.

[18] Notons que nous utilisons les termes « haut niveau » et « bas niveau » car nous faisons le postulat que le processus de perception de la parole est régit par l'intervention d'unités abstraites (voir e.g., Goldinger, 1998; Johnson, 2006; Pisoni & Levi, 2007, pour des hypothèses alternatives).

2.2. LE PROCESSUS DE RECONNAISSANCE DE MOTS

2.2.1. Problématique liée à la nature du signal acoustique de parole

La reconnaissance de mots est une activité nécessaire et essentielle à une communication réussie. Ainsi, ce processus est au cœur de la compréhension du langage oral et constitue de ce fait un sujet de recherche central en psycholinguistique. Les premiers travaux effectués sur la perception de la parole sont rapidement arrivés à dégager deux problématiques principales de ce phénomène.

La première vient du fait que le signal acoustique de parole est *continu* alors que tout individu adulte en a une perception *discrète*. Ainsi, lorsque nous entendons une phrase, nous la percevons comme composée par différentes unités distinctes les unes des autres (i.e., en mots). Cependant, l'onde sonore correspondant à cette perception ne dispose pas de frontière évidente (i.e., période de silence) entre les mots. Par conséquent, un auditeur se doit de *segmenter* la chaîne parlée afin de percevoir cette onde sonore comme composée par un ensemble d'unités discètes, plutôt que comme un seul et unique signal (voir e.g., Dumay, 2006; Shoemaker, 2009, pour des revues à ce sujet). Reconnaître un mot lorsque celui-ci est présenté dans le contexte d'autres mots constitue donc une activité complexe, bien qu'elle soit à première vue effectuée de manière automatique et sans effort particulier par un auditeur adulte écoutant un signal de parole prononcé dans sa langue maternelle (voir e.g., Shoemaker, 2009, pour des travaux effectués sur des apprenants d'une seconde langue). Néanmoins, l'objectif de cette thèse est d'étudier le processus de reconnaissance de mots *isolés*. Le problème de la segmentation lexicale (segmentation en mots) ne s'applique donc pas directement au travail présenté ici.

Un autre « challenge » perceptif auquel un auditeur est confronté vient du fait que lorsqu'un mot est prononcé, un nombre infini de productions acoustiques lui correspondent. En effet, l'onde sonore issue de la mise en action des différents articulateurs va être considérablement *variable* du point de vue du signal, alors que celle-ci va correspondre, sur le plan perceptif, à une seule et unique entité. En d'autres termes, il y a autant de versions acoustiques d'un mot que de réalisations. Il est possible de distinguer trois grandes sources de variabilité du signal acoustique, pour la perception d'un mot isolé. Premièrement, les conditions « physiques » (réverbération de l'onde sonore sur les parois d'une pièce, bruit environnant, caractéristiques d'un microphone et d'un téléphone, etc.) vont venir perturber la propagation de l'onde sonore consécutive à la mise en action des différents articulateurs de notre conduit vocal. Ensuite, une grande variabilité intra-locuteur (e.g., coarticulation de

certains segments, intonation de la voix, volume sonore, débit de parole, etc.) va rendre le signal acoustique de parole pour un même mot considérablement différent d'une production à l'autre. Enfin, une variabilité inter-locuteurs (e.g., forme et taille du conduit vocal différentes, timbre de la voix, présence d'accents, etc.) va également être à l'origine de nombreuses variations de cette onde sonore. Le constat de cette non-invariance acoustique a donc généré la seconde problématique à laquelle les chercheurs étudiant la perception de la parole tentent de répondre : comment un auditeur est-il capable de reconnaître un mot sur la base d'un signal acoustique aussi variable ? En réponse à cette question, il est généralement supposé qu'un auditeur dispose d'une (ou plusieurs) *représentation(s) abstraite(s)* stockée(s) en mémoire pour chaque mot qu'il connaît. Chacune de ces représentations (ou unités lexicales) serait constituée par un ensemble d'*invariants* permettant de reconnaître un signal de parole malgré ses variations acoustiques. L'objectif de la section suivante consiste à décrire ces représentations abstraites et à formuler les questions qui découlent d'un tel postulat.

2.2.2. Notion de lexique mental

En psycholinguistique, le *lexique mental* (Treisman, 1960, cité par Spinelli & Ferrand, 2005) désigne l'ensemble des représentations de mots connus par un individu. Issu de la métaphore du dictionnaire, il s'agit d'un système hautement organisé, contenant des unités reliées entre elles en fonction de leur similarité phonologique, orthographique, sémantique, syntaxique, etc. Dans cette perspective, la reconnaissance des mots parlés s'effectue en comparant l'entrée sensorielle (i.e., ce que l'on voit et/ou entend) et les représentations de ces mots en mémoire. En d'autres termes, il est possible de considérer ce processus comme résultant d'un appariement entre l'information sensorielle extraite du signal acoustique et les représentations lexicales mémorisées. Par « accès au lexique », on décrit l'opération par laquelle un signal acoustique va venir activer la ou les représentations lexicales correspondantes et permettre de reconnaître un mot.

C'est selon cette hypothèse abstractionniste que le travail détaillé dans ce manuscrit a été effectué. Dans les prochaines sections, nous allons examiner plusieurs questions primodiales pour l'étude du processus de reconnaissance de mots parlés, soulevées par le choix d'un tel cadre théorique.

2.2.3. Format des représentations permettant d'accéder au lexique

Lors du processus de reconnaissance de mots en modalité auditive, une des premières étapes consiste à extraire du signal acoustique les informations nécessaires pour accéder au lexique. Selon McQueen (McQueen, 2007), ces dernières peuvent être classées en deux grandes catégories. Alors que les informations *segmentales* désignent l'ensemble des caractéristiques du signal acoustique permettant de distinguer des sons de parole entre eux, les informations *suprasegmentales* font référence aux caractéristiques prosodiques des mots (e.g., l'intonation, l'accent tonique, la durée de prononciation de certains phonèmes). Dans ce manuscrit, nous nous intéresserons uniquement au rôle des informations segmentales dans le processus d'accès au lexique. Nous renvoyons le lecteur à d'autres travaux pour plus de détails à propos du rôle joué par les informations suprasegmentales (e.g., Cutler, Dahan, & Van Donselaar, 1997; Cutler & Norris, 2002; McQueen, 2007). La seconde étape du processus de reconnaissance de mots va consister à apparier ces informations fraîchement extraites du signal acoustique à des représentations stockées en mémoire. Or, pour accéder au lexique, la majorité des modèles actuels de perception de la parole supposent qu'il existe un niveau *pré-lexical* (e.g., McClelland & Elman, 1986; McQueen, 2007; Vitevitch & Luce, 1998) situé entre le lexique mental et l'entrée auditive, permettant d'effectuer ce contact. Le niveau pré-lexical désigne un ensemble de représentations plus petites que le mot (phonétiques, phonémiques, syllabiques, etc.) permettant une étape de décodage intermédiaire entre la stimulation sensorielle et les représentations lexicales. En plus de ce rôle de médiateur, plusieurs auteurs s'accordent à dire que ce niveau pré-lexical permettrait de faire face, conjointement avec le niveau lexical, au problème de l'invariance décrit ci-dessus (e.g., McQueen, 2005). Notons qu'il permet également aux modèles d'accès au lexique de rendre compte du fait qu'un individu est capable de percevoir et d'effectuer un certain nombre de traitements sur des signaux de parole qu'il n'a jamais rencontrés (e.g., pseudo-mots, nouveaux mots, mots prononcés dans une langue étrangère, etc.) et dont de fait il ne possède aucune représentation lexicale en mémoire. Dans la littérature, le format de ces représentations pré-lexicales est toujours l'objet de nombreuses controverses.

Parmi les unités fonctionnelles proposées, le phonème est l'une des représentations qui a été le plus souvent considérée à l'origine du contact avec le lexique (e.g., McClelland & Elman, 1986). Cependant, plusieurs années de recherches effectuées dans le domaine de l'acoustique et de la phonétique ont montré que le signal acoustique de parole ne constitue pas un enchaînement séquentiel et ordonné de phonèmes. En effet, notamment du fait de la

coarticulation, aucune correspondance directe ne peut être établie entre segment acoustique et un seul et unique phonème. Néanmoins, en dépit d'une absence de réalité *acoustique* de cette unité, le phonème constitue néanmoins une réalité *perceptive* pour tout auditeur. C'est pour cette raison que cette représentation est toujours proposée comme unité fonctionnelle permettant d'accéder au lexique dans certains modèles (e.g., TRACE, McClelland & Elman, 1986).

Le trait phonétique (e.g., Marslen-Wilson & Warren, 1994) a également été proposé comme pouvant jouer ce rôle (cf. section 2.3.1.2, modèle de la Cohorte II, Marslen-Wilson, 1987; Marslen-Wilson, 1990). Ainsi, dans une étude effectuée en anglais, Marslen-Wilson et Warren (1994) réfutent la nécessité d'un niveau de décodage phonémique situé entre le niveau des traits phonétiques et les représentations lexicales, tel que postulé par le modèle TRACE (McClelland & Elman, 1986, voir section 2.3.2 pour une présentation complète de ce modèle). Dans ce travail, les auteurs ont utilisé la technique de *cross-splicing*[19]. Cette méthode consiste à juxtaposer une portion du signal acoustique d'un énoncé de parole A avec un autre énoncé de parole B. Elle permet notamment de manipuler les caractéristiques subphonémiques d'un signal de parole tout en conservant les mêmes phonèmes. Ainsi, le *cross-splicing* de la consonne initiale et de la voyelle issues de /dʒɔb/, « job », travail, avec le phonème consonantique final issu de /dʒɔg/, « jog », footing, s'entend « jog » mais contient dans la portion vocalique, du fait de la coarticulation, une information acoustique relative à la place d'articulation d'un /b/. Les auteurs ont mesuré l'impact de cette discordance phonétique sur la capacité des participants à identifier chaque stimulus comme étant ou n'étant pas un mot (i.e., tâche de décision lexicale). Leurs résultats montrent que les participants avaient plus de difficultés à reconnaître un stimulus lorsque celui-ci présentait une discordance phonétique (condition ci-dessus), plutôt que lorsque l'ensemble des indices phonétiques était concordants entre eux (condition contrôle). Cette différence montrent qu'une discordance phonétique, en dehors de toute discordance phonémique, influence le processus d'accès au lexique. Ainsi, ces résultats indiquent que les représentations phonétiques peuvent constituer des unités fonctionnelles permettant de contacter *directement* le lexique (voir e.g., Goldinger, Luce, & Pisoni, 1989, pour des résultats contradictoires). Cela suggère également que les informations phonétiques sont intégrées directement au niveau lexical et pas à un niveau intermédiaire entre ces deux types de représentations (i.e., niveau phonémique). De ce fait ces auteurs réfutent l'hypothèse selon laquelle le phonème

[19] Littéralement : « raccord croisé »

constituerait une unité fonctionnelle privilégiée pour l'accès au lexique (voir e.g., McQueen, Norris, & Cutler, 1999; Stevens, 2002, pour des hypothèses similaires).

D'autres propositions ont également été effectuées : Klatt (1979), cité par Klatt (Klatt, 1989) suppose par exemple dans son modèle (LAFS, « Lexical Access From Spectra ») l'utilisation de gabarits spectraux comme représentation intermédiaire d'appariement entre le lexique et le stimulus d'entrée. La syllabe a également été proposée comme représentation abstraite permettant d'accéder au lexique (e.g., Mehler, Dommergues, Frauenfelder, & Segui, 1981; Mehler, Dupoux, & Segui, 1990; Segui, Dupoux, & Mehler, 1990). Mehler et al. (1981) ont pour cela utilisé des paires de mots partageant leurs trois phonèmes initiaux mais différant par la structure syllabique de leur première syllabe (e.g., « **pa**.**l**ace[20] » - « **pal**.mier »). Les participants avaient pour consigne de détecter le plus rapidement possible dans ces stimuli une cible CV (e.g., /pa/), ou CVC (e.g., /pal/). Ainsi, ces dernières pouvaient soit partager la première syllabe avec le mot porteur (e.g., /pa/ - /pa.las/, « palace ») soit ne pas la partager (e.g., /pa/ - /pal.mie/, « palmier »). Leurs résultats montrent clairement une interaction entre la structure syllabique des cibles et celles des mots porteurs. Celle-ci indique que les participants étaient plus rapides pour détecter la cible lorsqu'elle correspondait à la première syllabe du mot porteur (e.g., « pa » dans « palace » et « pal » dans « palmier »). Or, dans l'hypothèse que le signal acoustique d'un mot est tout d'abord décodé en phonèmes avant de contacter le lexique, aucune différence n'aurait dû être obtenue quel que soit l'alignement syllabique de l'amorce avec le mot porteur. En effet, rappelons que les trois premiers phonèmes de chaque élément de la paire de mots porteurs étaient identiques, seule la structure syllabique de cette portion était différente. Ces résultats suggèrent donc que la syllabe contraint l'accès au lexique chez des auditeurs de langue maternelle française (voir e.g., Cutler, Mehler, Norris, & Segui, 1986; Sebastián-gallés, Dupoux, Segui, & Mehler, 1992, pour une modulation de cet effet en fonction de la langue). Cette idée est en accord avec l'hypothèse que la syllabe pourrait constituer une unité à part entière permettant de contacter les représentations lexicales.

En conclusion, il semblerait que le trait phonétique (e.g., Marslen-Wilson & Warren, 1994), le phonème (e.g., McClelland & Elman, 1986) et la syllabe (e.g., Mehler, et al., 1981) puissent constituer des unités fonctionnelles pour accéder aux représentations lexicales. Ainsi, cette question est toujours non résolue dans la littérature et constitue un point sur lequel les modèles psycholinguistiques actuels divergent. Alors que certains supposent que

[20] Les lettres en gras correspondent aux phonèmes partagés entre les paires alors que le point indique les frontières syllabiques des deux items.

c'est le trait phonétique (e.g., Cohorte II, Marslen-Wilson, 1987, 1990) ou le phonème (e.g., TRACE, McClelland & Elman, 1986, Shortlist, Norris, 1994) qui constitue l'unité privilégiée permettant d'accéder aux représentations lexicales, d'autres postulent que c'est la syllabe (e.g., SARAH, Mehler, et al., 1990) qui permet d'effectuer ce contact. Remarquons cependant que dans les travaux que nous venons de passer en revue, la positions dans le mot des diverses unités étudiées était différente. En d'autre termes, cela signifie que l'unité permettant d'accéder au lexique a été interrogée à différents moments dans le mot. L'objectif de la prochaine section consiste justement à étudier le décours temporel de ce processus.

2.2.4. Décours temporel de l'accès au lexique

L'onde sonore est, par définition, distribuée dans le temps. De ce fait, les différentes informations auditives contenues dans un signal acoutique de parole ne vont pas être disponibles au même moment dans le langage oral. Si au niveau acoustique, le signal de parole est séquentiel, en est-il de même pour son traitement ? La parole est-elle décodée au fur et à mesure, de manière continue dans le temps ? Et si oui, à partir de quelle quantité d'information disponible ? Toujours dans cette hypothèse, l'information partielle qui en serait extraite est-elle directement exploitée pour contacter les représentations lexicales ou est-elle seulement décodée à un niveau pré-lexical ?

Parmi un grand nombre d'études ayant exploré ces problématiques, celles effectuées par Marslen-Wilson et Warren à la fin des années 1980 (e.g., Warren & Marslen-Wilson, 1987, 1988) ont apporté des éléments de réponse à ces questions. Dans une étude effectuée en anglais, Warren et Marslen-Wilson (1988) ont utilisé des paires de mots monosyllabiques ne différant que par un seul trait phonétique : la place d'articulation (e.g., /slɔp/, « slop », pâtée, vs. /slɔt/, « slot », fente). Chaque élément de ces paires était équivalent en termes de fréquence dans le langage oral. Ces stimuli étaient présentés à l'aide d'un paradigme de *gating*. Les participants avaient pour tâche d'identifier le mot qu'ils étaient censés avoir perçu. Les résultats de cette étude sont illustrés dans la Figure 15.

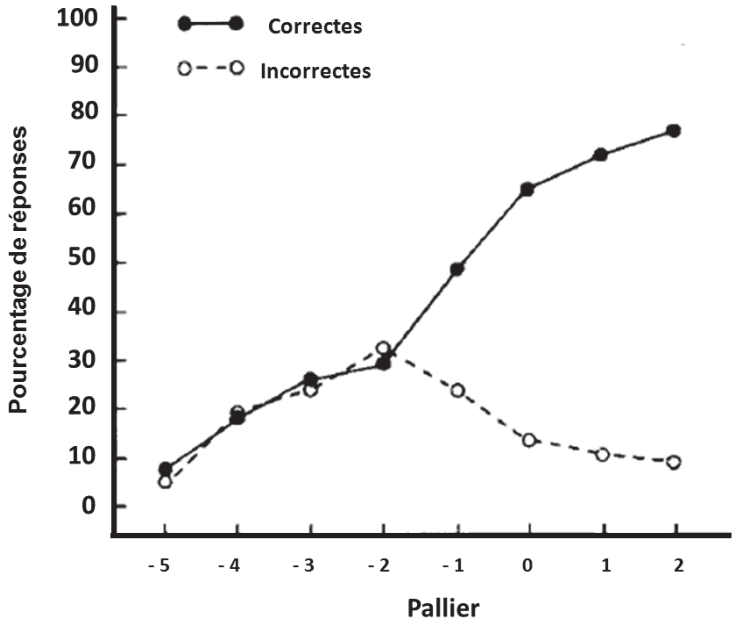


Figure 15. *Pourcentages d'identifications correctes (e.g., « slop ») et incorrectes (e.g., « slot ») en fonction de la quantité d'information dévoilée. L'intervalle temporel entre chaque pallier est de 25 ms. Le pallier 0 correspond à la fin acoustique de la voyelle. (Adapté de Warren & Marslen-Wilson, 1988).*

L'examen de leurs résultats montre tout d'abord que pour les les 100 premières ms du stimulus, autant de réponses correctes (e.g., « slop ») qu'incorrectes (e.g., « slot ») sont effectuées pour identifier le mot en question (e.g., « slop »). Cela indique qu'aucune préférence n'existait pour un élément de la paire en particulier. Ensuite, pour les 50 ms suivantes, on observe une augmentation des identifications correctes accompagnée d'une diminution des réponses incorrectes. Ce changement est observé pendant la production de la voyelle, alors que l'ensemble des informations sur la consonne finale n'est pas encore disponible dans le stimulus (i.e., entre 25 et 50 ms avant la fermeture du conduit vocal pour le phonème final). Ainsi, un auditeur serait capable d'extraire du signal de parole différentes informations pour contacter le lexique *avant* la fin du mot. Plus précisément, cette étude suggère que le décodage du signal de parole s'effectuerait en *continu* pour accéder au lexique, chaque information partielle nouvellement disponible étant directement traitée pour moduler les choix lexicaux de l'auditeur (voir e.g., Dahan, Magnuson, & Tanenhaus, 2001, pour des arguments similaires obtenues avec des méthodes d'enregistrement des mouvements occulaires). Dans ce travail, c'est l'information acoustique relative à la place d'articulation (e.g., bilabiale) du phonème consonantique final (e.g., /p/) qui du fait de la coarticulation, est présente dans la portion vocalique du stimulus qui semble avoir été traitée par les participants pour anticiper l'identité du mot. Des résultats similaires ont également été retrouvés dans cette même étude pour d'autres traits phonétiques consonantiques (e.g., le voisement).

Par conséquent, l'examen de cette première étude indique qu'un mot peut être identifié *avant* que son signal acoustique ne soit terminé. Il a été estimé que le contact avec les représentations lexicales se déclencherait entre 150 (Tyler & Wessels, 1983) et 200 ms (Marslen-Wilson, 1987) après son début acoustique. Certains signaux de parole ne pouvant être reconnus qu'à leur toute fin (e.g., Grosjean, 1980), nous attirons l'attention du lecteur sur le fait que l'obtention de ces chiffres doivent être considérés comme une valeur moyenne calculée à partir de résultats issus d'études différentes. Cependant, cette valeur constitue globalement la durée d'une syllabe. En dehors de toute considération relative au format d'unité (e.g., syllabe, phonème, trait phonétique, etc.) permettant de contacter le lexique (voir section 2.2.3 pour plus de détails à ce sujet), cette fenêtre temporelle semble également indiquer que le *début* de mot pourrait (en comparaison avec la fin du mot) constituer un rôle important dans le processus d'accès aux représentations lexicales. Plusieurs travaux indiquent en effet que le début de mot jouerait un rôle plus important que sa fin dans le processus d'accès au lexique (e.g., Luce & Lyons, 1999; Marslen-Wilson & Zwitserlood, 1989; Spinelli, 1999; Spinelli, Segui, & Radeau, 2001; Zwitserlood, 1989). Pour étudier cette question, la majorité des études citées ont utilisé des paradigmes d'*amorçage intermodaux*. Le paradigme d'amorçage est basé sur l'idée que la reconnaissance d'un mot (i.e., appelé « amorce ») peut influencer (en termes de facilitation ou d'inhibition) le traitement d'un autre mot (i.e., appelé « cible ») si une relation (sémantique, phonologique, etc.) existe entre eux. La tâche demandée aux participants est le plus souvent effectuée sur la cible. L'effet d'amorçage est mesuré en comparant les performances dans une condition où l'amorce et la cible entretiennent une relation (condition reliée) avec une condition où ces dernières n'entretiennent aucune relation entre elles (condition non reliée). Les modèles décrivant l'accès au lexique postulent que lorsqu'un mot est reconnu, la représentation de ce dernier est activée en mémoire. Les effets d'amorçages sont expliqués par le fait qu'en condition reliée, percevoir l'amorce va permettre d'activer un certain nombre de représentations en mémoire et notamment celle de la cible. Ainsi, l'activation résiduelle liée à la présentation de l'amorce va permettre de faciliter la reconnaissance ultérieure de la cible par rapport à une condition non reliée, où l'amorce n'est pas censée avoir activé la représentation de la cible. Le paradigme d'amorçage intermodal (cf. Figure 16) consiste (en opposition avec un paradigme d'amorçage intramodal) à présenter l'amorce et la cible dans une modalité sensorielle différente (e.g., amorce auditive, cible visuelle).

La logique sous-tendant l'utilisation de tels paradigmes est qu'un effet d'amorçage intramodal peut être uniquement lié à l'intervention de mécanismes spécifiques à la modalité (i.e., pré-lexicaux). Un effet d'amorçage intermodal nécessite que deux sources

d'informations soient décodées à un niveau non spécifique à la modalité, donc plus abstrait, voire amodal (i.e., lexical). Si un effet d'amorçage est uniquement présent en intramodal mais pas intermodal, on en conclut que cet effet est plutôt lié à l'intervention de processus pré-lexicaux plutôt que lexicaux (voir, Grosjean & Frauenfelder, 1996, pour une revue complète sur les différents paradigmes utilisés pour étudier le processus de reconnaissance de mots).

Pour étudier l'importance de début de mot dans l'accès au lexique, Spinelli et al. (2001) ont utilisé un paradigme d'amorçage phonologique partiel en situations inter et intramodales (voir Dufour, 2003; Spinelli, 1999, pour des revues sur l'amorçage phonologique).

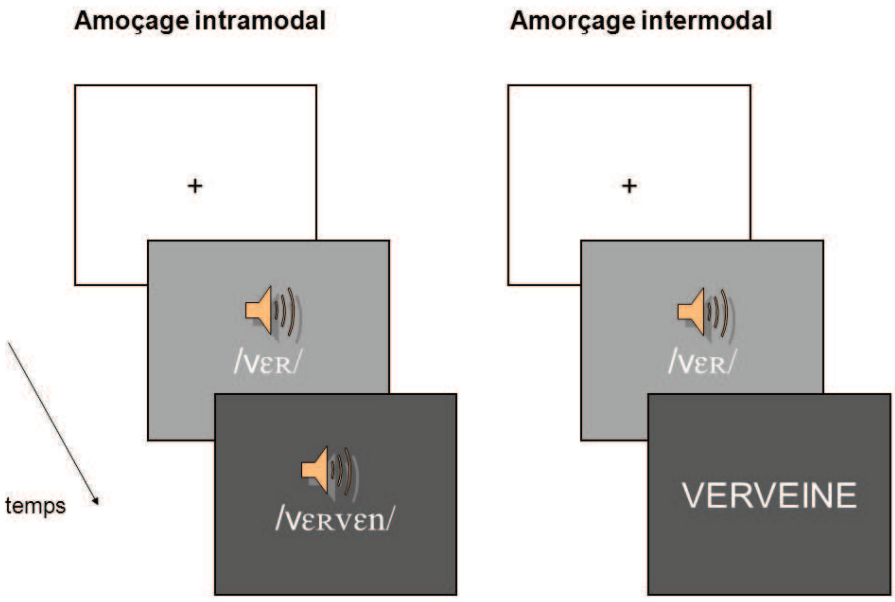


Figure 16. *Représentation schématique des paradigmes d'amorçage intra et intermodal dans le cadre de l'étude de Spinelli et al. (2001). Dans le cas de l'amorçage intra-modal, l'amorce (en gris clair) est présentée dans la même modalité (e.g., auditive) que la cible (en gris foncé). Pour le paradigme d'amorçage intermodal, l'amorce et la cible sont présentées une modalité différente (auditive vs. visuelle écrite).*

L'amorce pouvait soit partager le même début avec la cible (condition de recouvrement initial, e.g., /vɛʁ/, « ver » → /vɛʁvɛn/, « verveine ») soit la même fin (condition de recouvrement final, e.g., /vɛn/, « veine » → /vɛʁvɛn/, « verveine »). Dans une condition contrôle, l'amorce ne partageait aucune relation (phonologique, sémantique, etc.) avec la cible (e.g., /kwɛ̃/, « coin » → /vɛʁvɛn/, « verveine »). L'amorce était toujours présentée en modalité auditive. En revanche, la cible était soit présentée en modalité auditive (amorçage intramodal), soit à l'écrit en lettres majuscules, c'est-à-dire en modalité visuelle (e.g., amorçage intermodal). La tâche demandée aux participants était une tâche de décision

lexicale[21]. L'examen des temps de réponse observés pour la situation intramodale indique premièrement qu'un effet de facilitation est observé pour la condition de recouvrement final et initial, par rapport à la condition contrôle. Cela indique que le signal acoustique contenu dans la première syllabe (ou dans les trois premiers phonèmes de l'amorce) constitue une information suffisante pour faciliter le traitement ultérieur de la cible. Cependant, en situation intermodale, l'effet de facilitation est répliqué pour la condition de recouvrement initial, mais pas pour la condition de recouvrement final. Conformément à la logique des études intermodales, ces résultats suggèrent que l'effet de facilitation observé en recouvrement final (e.g., Radeau, Morais, & Segui, 1995; Slowiaczek, Nusbaum, & Pisoni, 1987) relève plutôt d'un locus pré-lexical que lexical (voir aussi e.g., Dumay et al., 2001, pour une conclusion similaire). En revanche, dans cette étude, l'effet d'amorçage obtenu en recouvrement inital est obtenu en situation intra et intermodale. Cela indique que l'effet de facilitation observé serait plutôt lié à l'intervention de mécanismes abstraits, lexicaux. En d'autres termes, cela suggère que le signal acoustique de l'amorce (e.g., « ver ») constitue une information suffisante pour activer la représentation lexicale de la cible (e.g., « verveine »). Ainsi, les trois premiers phonèmes (ou la première syllabe) d'un mot permettent de contacter le niveau lexical. Au contraire, la présentation des trois derniers phonèmes (ou dernière syllabe) d'un mot ne semble pas de faciliter le processus de reconnaissance de mots per se, mais uniquement le décodage de la parole à un niveau pré-lexical. Ce résultat souligne donc l'importance du début de mot dans l'accès au lexique.

D'autres travaux ont également testé cette hypothèse en évaluant si une discordance située à l'initiale d'un mot pouvait gêner voire bloquer l'accès à sa représentation lexicale (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Connine, Blasko, & Titone, 1993; Frauenfelder, Scholten, & Content, 2001; Marslen-Wilson & Zwitserlood, 1989). En utilisant un pardigme d'amorçage sémantique, Marslen-Wilson et Zwitserlood (1989) ont pour leur part obtenu des résultats semblant indiquer que le partage du début des mots est nécessaire à l'activation des candidats lexicaux. Dans une étude effectuée en néerlandais, ils ont montré que la présentation auditive de « honing », miel, facilite le traitement du mot écrit « BIJ », abeille, associé sémantique de « honing ». Néanmoins, entendre le mot « woning », accueil, ou le pseudo-mot « foning » qui ne diffèrent du mot « honing » que par leur premier phonème ne facilitent pas le traitement du mot-cible « BIJ ». Ainsi, ces résultats suggèrent qu'une discordance d'un seul phonème située à l'initiale d'un mot constitue

[21] Pour les besoins de cette tâche, des pseudo-mots cibles ont également été utilisés. Ces derniers étaient également répartis dans les trois conditions. Aucune analyse sur les performances obtenues sur ces items n'a été effectuée.

suffisamment d'information pour bloquer l'accès à sa représentation lexicale, même lorsque cette modification donne lieu à la perception d'un pseudo-mot. En utilisant un paradigme similaire, Connine et al. (1993) ont également manipulé la distance phonologique entre les phonèmes divergeants. Ils ont rapporté une facilitation du traitement sémantique par la présentation du pseudo-mot amorce (e.g., /pænər/, « panner ») alors que celui-ci différait de un ou deux traits phonétiques avec le phonème initial du mot source (e.g., /bænər/, « banner », bannière). Ces résultats semblent donc indiquer que l'accès au lexique n'est pas complètement bloqué lors d'une discordance faible entre l'entrée sensorielle et sa représentation (voir Frauenfelder, et al., 2001, pour des résultats comparables). Toutefois, dans l'ensemble de ces études, les effets d'amorçage facilitateurs entre des stimuli qui n'ont pas le même début ont été uniquement obtenus lorsque la modification du mot source donnait lieu à un pseudo-mot plutôt qu'à un mot réel. Les différents travaux que nous venons de passer en revue suggèrent qu'aucune facilitation n'est observée lorsque cette manipulation donne lieu au contraire à un mot réel (e.g., Marslen-Wilson & Zwitserlood, 1989). En effet, ces auteurs (Marslen-Wilson, Moss, & Van Halen, 1996) font l'hypothèse que le contact initial avec le lexique exclurait dans un premier temps les mots qui n'ont pas le même début que celui entendu. Ils supposent que lorsqu'un auditeur parvient à reconnaître un mot A malgré une discordance située à l'initiale (et pouvant donner lieu à la perception d'un mot B) c'est par l'intervention de processus qui opèreraient *après* la première phase de contact avec le lexique mental.

L'étude effectuée par Allopenna et al. (1998) a permis de montrer que la présence d'une discordance phonémique à l'initiale d'un mot A, donnant lieu à la perception d'un mot B pouvait néanmoins activer la représentation de A. Pour cela, ces auteurs ont mesuré les mouvements oculaires des participants lorsque ceux-ci entendaient des instructions leur indiquant de cliquer sur l'image correspondant au mot entendu. A chaque essai, quatre images différentes étaient présentées sur un écran d'ordinateur. Le nom de ces objets pouvait soit correspondre au mot entendu (image de référence, e.g.,/biːkər/, « beaker », gobelet), soit partager la même fin (distracteur « rime », e.g., /spiːkər/, « speaker », enceinte), soit commencer par le même début (distracteur « début », e.g., /biːtl/, « beetle », coccinelle), soit ne partager aucun phonème avec celui-ci (distracteur « contrôle », e.g., /kærɪdʒ/, « carriage », carrosse). Leurs résultats montrent qu'une plus grande proportion de fixations était accordée aux images rimant avec la cible (e.g., speaker) plutôt qu'aux images contrôles (e.g., « carriage »), avant même la fin acoustique du mot. Bien que ce pourcentage soit inférieur aux temps de fixation observés séparément pour le distracteur « début » et l'image

de référence, leurs données semblent indiquer qu'un appariement au sens strict entre l'information sensorielle initiale et sa représentation lexicale correspondante ne soit pas une condition nécessaire à son accès. Notons que ces conclusions sont en oppositions avec l'idée défendue par Marslen-Wilson et collègues (e.g., Marslen-Wilson et al., 1996).

En conclusion, l'examen de ces différentes études nous renseigne sur le fait que l'accès au lexique peut débuter avant la fin du signal acoustique de parole (e.g., Warren & Marslen-Wilson, 1988). L'information extraite de cette activité de décodage serait exploitée en temps réel par notre système perceptif et de manière continue (e.g., Dahan, et al., 2001), permettant à un auditeur de moduler ses choix lexicaux au fur et à mesure du décours temporel du signal (e.g., Warren & Marslen-Wilson, 1988). L'accès au lexique pouvant se déclencher après les 200 *premières* ms d'un mot (e.g., Marslen-Wilson, 1987), il semblerait que le début de mot joue (relativement à sa fin) un rôle important dans ce processus (e.g., Marslen-Wilson & Zwitserlood, 1989; Spinelli, et al., 2001), bien qu'un appariement phonologique strict entre le début du stimulus auditif et la représentation phonologique lui correspondant ne soit pas forcément nécessaire (e.g., Allopenna, et al., 1998; Connine, et al., 1993). Ainsi, ce point est toujours sujet à controverse dans la littérature et oppose les modèles d'accès au lexique accordant un statut privilégié au début de mot (e.g., Cohorte II, Marslen-Wilson, 1987, 1990) à ceux qui ne postulent pas un tel statut (e.g., NAM, Luce & Pisoni, 1998). L'objectif de la prochaine section consiste à étudier l'influence des représentations lexicales sur le décodage d'un signal acoustique de parole.

2.2.5. Influence de l'information lexicale sur la perception de la parole

A notre connaissance, la première étude à avoir mis en évidence l'influence de l'information lexicale sur la perception de la parole a été effectuée en anglais par Rubin et collègues (Rubin, Turvey, & Van Gelder, 1976). L'objectif des auteurs était d'examiner l'implication du niveau lexical sur le traitement des phonèmes. Les auteurs ont par conséquent proposé une tâche de détection de phonèmes consonantiques dans des mots (e.g., /b/ dans /**b**ɪt[22]/, morceau) ou des pseudo-mots monosyllabiques (e.g., /b/ dans /**b**ɪp/). Cette distinction avait pour but de distinguer les mécanismes opérant uniquement à un niveau lexical, pouvant être observés seulement sur les mots (les pseudo-mots ne disposant pas de représentation lexicale en mémoire) de ceux opérant à un niveau pré-lexical, pouvant être observés à la fois sur les mots et sur les pseudo-mots. Notons que dans ce travail, des

[22] La lettre en gras indique le phonème à détecter

pseudo-mots ont été préférentiellement utilisés par rapport à des non-mots afin de manipuler des items non lexicaux tout en respectant les contraintes phonotactiques de l'anglais. Ainsi, les prédictions effectuées par les auteurs étaient que si l'information lexicale influence le processus de détection des phonèmes-cibles, les performances obtenues à cette tâche devraient varier en fonction du statut lexical des items contenant la cible à détecter. Les résultats montrent qu'effectivement, les participants étaient plus rapides pour détecter le phonème-cible lorsque celui-ci est situé dans un mot plutôt que dans un pseudo-mot. Cet effet de supériorité du mot indique que le contexte lexical influence le processus de détection de phonèmes ; cet effet a été répliqué plusieurs fois depuis (e.g., Cutler, Mehler, Norris, & Segui, 1987, Expérience 1, 4 et 6 ; Frauenfelder, Segui, & Dijkstra, 1990, Expérience 1), (voir Connine & Titone, 1996, pour une revue). Par exemple, les travaux effectués en néerlandais par Frauenfelder et al. (1990) se sont intéressés à cet effet dans des mots et des pseudo-mots plurisyllabiques en faisant varier la position du phonème-cible à détecter en fonction des points d'unicité des mots et non-mots porteurs de celui-ci. Pour les mots, ce que nous avons désigné comme point d'unicité (« uniqueness point ») fait référence au moment (en partant du début du mot) pour lequel celui-ci devient l'unique candidat possible. Pour les pseudo-mots, ce terme (« nonword point ») désigne ici le moment (en partant de son début) à partir duquel l'item est forcément un pseudo-mot. Ainsi, le phonème-cible (e.g., /p/) pouvait se situer à l'initiale, avant le point d'unicité (e.g., « **p**agina », page vs. « **p**afime ») ou en position finale, après le point d'unicité (e.g., « bioscoo**p** », cinéma vs. « deoftoo**p** »). Leurs résultats ont mis en évidence que le point d'unicité joue un rôle important sur l'influence du niveau lexical : en effet ces derniers n'ont trouvé un effet de supériorité du mot que lorsque le phonème était situé après le point d'unicité, mais pas avant. En d'autres termes, les auteurs ont mis en évidence que les participants étaient plus rapides pour détecter le phonème-cible dans un mot plutôt que dans un pseudo-mot lorsque celui-ci était situé à la fin de l'item, mais pas au début. Les auteurs en ont conclu que le niveau lexical influence la perception d'unités pré-lexicales non ambigües majoritairement après son point d'unicité, c'est-à-dire lorsque le traitement lexical est bien avancé.

Une autre étude princeps dans ce domaine est celle effectuée en anglais par Ganong (Ganong, 1980). Cet auteur avait pour ambition de tester l'influence du niveau lexical sur la catégorisation de phonèmes consonnantiques occlusifs (e.g., /d/ et /t/). Un paradigme classiquement utilisé pour étudier le phénomène de catégorisation phonétique consiste à manipuler sur un continuum une variable acoustique (e.g., le délai de voisement ou *VOT*) permettant de différencier deux phonèmes entre eux (e.g., /d/ vs. /t/). Le délai de voisement, couramment appelé « Voice Onset Time », *VOT*, correspond à l'intervalle entre

le relâchement de l'air bloqué par le conduit vocal avant l'ouverture de la bouche et le déclenchement des vibrations des cordes vocales. Ainsi, lorsque cet intervalle est petit (i.e., lorsque la durée de vibration des cordes vocales est longue) la consonne est perçue comme étant voisée (e.g., /d/). Lorsque celui-ci est grand (i.e., lorsque la durée de vibration des cordes vocales est courte), la consonne est perçue comme non voisée (e.g., /t/). Lorsque l'on demande à des participants de langue maternelle anglaise d'identifier un phonème placé sur un continuum de voisement /da-ta/, la frontière entre le /d/ et le /t/ correspond environ à 35 ms de *VOT*. Les résultats montrent que généralement, un auditeur anglais va systématiquement (1) identifier un /d/ pour des valeurs inférieures à 30 ms de *VOT* et (2) identifier un /t/ pour des valeurs de *VOT* supérieures à 40 ms. Ainsi, seule la petite portion de valeurs de *VOT* comprises entre 30 et 40 ms correspondent à un percept ambigu. Ce phénomène de perception catégorielle renvoie à l'observation d'un basculement brutal de la *perception* de /d/ à /t/ pour un changement graduel d'une variable acoustique du signal. Dans l'expérience effectuée par Ganong (1980), un paradigme similaire a été utilisé excepté que le phonème à identifier était présenté dans un continuum de *VOT* formé par un pseudo-mot à une extrémité (e.g., /**t**aʃ[23]/ ou /**d**ask/) et par un mot à l'autre extrémité (e.g., /**d**aʃ/, « **d**ash », tiret, ou /**t**ask/, « **t**ask », tâche).

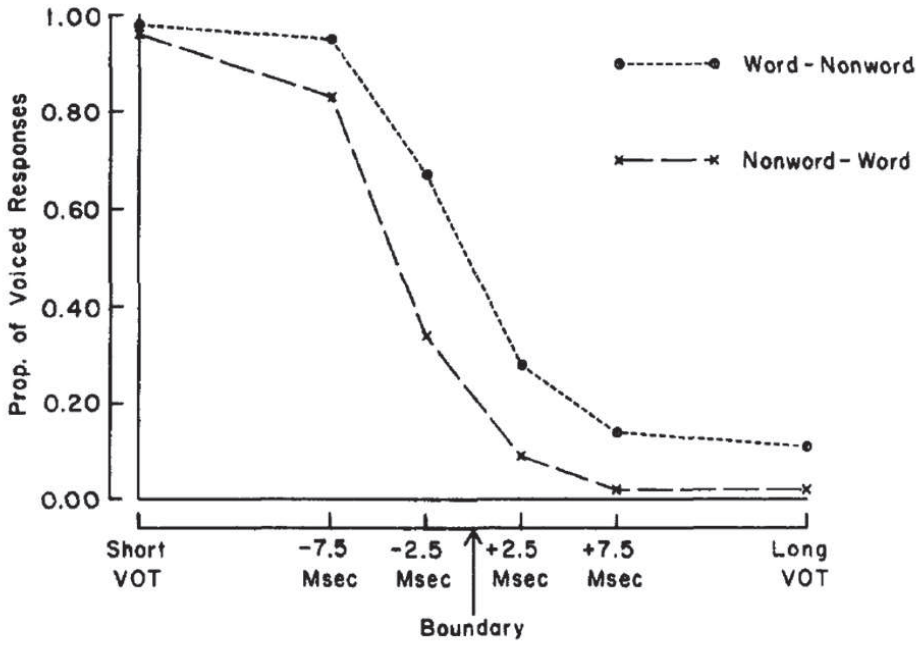


Figure 17. *Pourcentage d'identification de phonèmes voisés en fonction de la durée du VOT. Les petits traits pointillés représentent le continuum où le phonème voisé est contenu dans le mot (e.g., /**d**aʃ-**t**aʃ/) alors que les traits pointillés les plus longs représentent le continuum où le phonème voisé est contenu dans le pseudo-mot (e.g., /**d**ask-**t**ask/). (Extrait de Ganong, 1980).*

[23] La lettre en gras indique le phonème dont le VOT est manipulé

Leurs résultats mettent en évidence un déplacement de la frontière catégorielle classiquement observée pour le continuum syllabique /da-ta/ en faveur du phonème contenu dans le mot (cf. Figure 17). Ainsi, les auteurs ont observé une influence importante de l'information lexicale sur le processus de catégorisation phonémique et ce surtout lorsque le phonème à identifier est complètement ambigu, c'est-à-dire lorsqu'il est situé au milieu du continuum (voir aussi, e.g., Connine & Clifton, 1987; Fox, 1984; Pitt, 1995, pour des résultats comparables). Ainsi, ce résultat indique, comme les précédents, une influence du niveau lexical sur le traitement des unités pré-lexicales. Cependant, l'inconvénient de cette tâche, est que contrairement au paradigme de détection de phonèmes, celle-ci est effectuée en temps différé par les participants et non en temps réel, c'est-à-dire après la présentation des stimuli et non lors de celle-ci. Cette tâche ne permet donc pas de mesurer les performances du participant au moment où celui-ci perçoit le stimulus mais seulement après. Le paradigme de Ganong dispose donc d'un biais méthodologique similaire à celui du *gating* (cf. section 1.3.3 du Chapitre 1) dans le sens où les performances des participants peuvent être contaminées par des biais de stratégies conscientes.

Enfin le paradigme de « restauration phonémique » a également servi à étudier l'influence du niveau lexical sur le décodage du signal de parole (e.g., Samuel, 1981, 1996). Cet effet a été mis en évidence pour la première fois par Warren (Warren, 1970). Le phénomène de restauration phonémique survient lorsqu'une portion d'un mot correspondant à un phonème est remplacé par du bruit. Il met en évidence qu'un mot est perçu comme intact dans cette condition, de la même manière que lorsque le bruit est ajouté (plutôt que substitué) au phonème. Ainsi, les participants réagissent comme si le bruit ne faisait que recouvrir le signal de parole alors que celui-ci est en réalité absent. On dit alors que le segment manquant dans le signal est automatiquement « restauré ». Samuel (1981) a montré que cet effet de restauration phonémique avait principalement lieu dans des mots (e.g., "pr**o**gress[24]", /pr**əʊ**grɛs/) plutôt que dans des pseudo-mots (e.g., "cr**o**gless", /kr**əʊ**glɛs/). En d'autres termes, les participants éprouvent plus de difficultés à distinguer les stimuli dont le segment est intact (phonème + bruit) des stimuli dans lesquels cette portion du signal a été supprimée (bruit) lorsqu'ils identifient des mots plutôt que lorsqu'ils traitent des stimuli qui n'ont pas de représentation spécifique dans le lexique. Cet effet indique, à l'instar des études précédentes, que le contexte lexical influence la perception du signal de parole, permettant de résoudre l'ambiguïté lorsqu'une portion du signal acoustique est détériorée par du bruit

[24] La lettre en gras indique le phonème manquant

ou absente. Notons toutefois que la tâche de restauration phonémique dispose des mêmes biais méthodologiques que le paradigme de Ganong.

Dans cette section, nous avons avancé plusieurs preuves expérimentales indiquant, à l'aide de paradigmes différents (tâche de détection de phonèmes, e.g., Frauenfelder et al., 1990 ; effet Ganong, e.g., Ganong, 1980 ; effet de restauration phonémique, e.g., Samuel, 1981) que l'information lexicale biaise le traitement de la parole à un niveau pré-lexical. Notons que ces paradigmes servent aujourd'hui d'outils permettant de mesurer l'implication du niveau lexical sur le décodage du signal de parole. Cependant, le processus à l'origine de cette influence reste sujet à débat dans la littérature. En effet, alors que certains modèles (e.g., TRACE, McClelland & Elman, 1986) supposent l'existence d'un mécanisme rétroactif du niveau lexical vers les niveaux pré-lexicaux pour expliquer ces effets, d'autres (e.g., Merge, Norris, McQueen, & Cutler, 2000) postulent que ce type d'influence résulterait plutôt de l'intervention de processus décisionnels (et non perceptifs). Ces différents mécanismes seront examinés plus en détails dans la section suivante, consacrée à la présentation de différents modèles cognitifs décrivant l'accès au lexique.

2.3. MODELES DECRIVANT L'ACCES AU LEXIQUE EN MODALITE AUDITIVE

L'objectif de cette section consiste à présenter différents modèles décrivant l'accès au lexique. La majorité d'entre eux (mais voir section 2.3.6) postulent l'existence d'un seul signal d'entrée : le signal acoustique. Dans cette section, nous avons choisi de ne présenter que six modèles de manière approfondie. Nous avons ainsi sélectionné ceux semblant le plus résister dans la littérature aujourd'hui. Nous renvoyons le lecteur à d'autres articles pour des explications concernant les modèles suivants : SARAH (Mehler, et al., 1990) ; « Lexical Access From Spectra » (Klatt, 1979, cité dans Klatt, 1989). Remarquons aussi que nous avons décidé d'insister uniquement sur les caractéristiques de ces modèles étant directement reliées à notre travail.

Avant de s'attarder sur les aspects spécifiques inhérents à chacun d'entre eux, signalons que l'ensemble de ces modèles postulent que les unités permettant d'apparier le signal de parole avec sa représentation en mémoire sont de nature phonologique. Egalement deux grands mécanismes sont communs à ces modèles : l'activation multiple de candidats et le phénomène de compétition (ou sélection) lexicale. En effet, ils postulent qu'une *multitude* de représentations lexicales (ou *candidats* lexicaux) compatibles avec au moins une portion de

l'entrée auditive vont être activées *simultanément* et de manière *automatique* par l'arrivée d'un signal acoustique de parole. Nicolas Dumay (Dumay, 2006) fait d'ailleurs remarquer à ce propos que l'hypothèse de cette activation simultanée est présente dans tous les modèles psycholinguistiques et ce depuis la version initiale de la théorie de la Cohorte (Marslen-Wilson & Welsh, 1978). Ensuite, l'ensemble des modèles présentés dans cette section postule également une phase de *compétition* ou de *sélection* (implémentées de manière différentes en fonction des architectures) entre les candidats lexicaux précédemment activés. Un mot est reconnu à l'issue de cette étape, lorsqu'il ne reste plus qu'un seul candidat activé.

2.3.1. Le modèle de la Cohorte

2.3.1.1. « Active Direct Access Model » ou modèle de la Cohorte version I (Marslen-Wilson & Welsh, 1978)

Historiquement, le modèle de la Cohorte version I (« Active Direct Acess Model »), est le premier à avoir été formalisé pour rendre compte de la perception de mots parlés (Marslen-Wilson & Welsh, 1978). En d'autres termes, il s'agit du premier modèle ayant décrit l'accès au lexique en tenant compte des caractéristiques inhérentes aux propriétés du signal acoustique, tous les autres ayant jusqu'à cette date concerné les mots écrits. Celui-ci suppose, à l'instar de ses concurrents, deux étapes successives pour la reconnaissance d'un mot. Dans ce modèle, la perception d'un signal acoustique de parole active, dans un premier temps, une multitude de candidats lexicaux stockés dans la mémoire de l'auditeur. Cet ensemble de représentations activées forme la *cohorte initiale*. Un des paramètres singuliers de ce modèle est qu'il postule que l'appariement entre le signal d'entrée et la représentation lexicale s'effectue par rapport au *début* de l'entrée acoustique. Ainsi, seuls les candidats partageant les mêmes phonèmes initiaux que le stimulus extérieur sont sélectionnés pour faire partie de la cohorte initiale. Par exemple, la séquence de phonèmes /ba/ permet de former une cohorte initiale dont l'ensemble des candidats commencent par /b/ (e.g., /bato/, « bateau », /balɛ/, « balai », etc.). Notons que dans cette phase, pour cette version du modèle, l'activation des candidats est supportée par un mécanisme binaire : soit il est activé, du fait d'un appariement phonémique *parfait* entre le début du signal et sa représentation, soit il n'est pas activé et ne fait pas partie de la cohorte initiale. Remarquons que l'alignement, permettant d'apparier les éléments du signal compatibles avec la ou les représentations stockées en mémoire est effectué grâce à la mise en jeu d'unités *phonémiques*.

En d'autres termes, le modèle de la cohorte I postule que c'est le *phonème* qui va servir d'unité fonctionnelle pour contacter le lexique.

Dans la deuxième étape, l'état d'activation de chacune des représentations de la cohorte initiale se modifie en fonction de la correspondance de ces dernières avec le signal. Cette correspondance est évaluée de « gauche à droite », c'est-à-dire du début vers la fin acoustique du mot. Le processus permettant d'assurer cette opération est, à l'instar de celui permettant de former la cohorte initiale, un mécanisme binaire en « tout ou rien ». De cette manière, ce modèle postule que le moindre défaut d'appariement entre l'entrée acoustique et une représentation lexicale va engendrer l'*exclusion* de cette représentation de la cohorte, c'est-à-dire de l'ensemble des candidats lexicaux. Ce processus permet donc de sélectionner l'unité lexicale correspondant à l'entrée acoustique en réduisant progressivement le nombre de candidats lexicaux de la cohorte, plus la quantité d'information disponible est importante. L'accès au lexique est donc considéré comme un processus *séquentiel*. Un mot est reconnu lorsqu'il ne reste plus qu'un seul candidat dans la cohorte. Remarquons ici que les mécanismes supposés sont uniquement *ascendants* ou « bottom-up ». Ainsi la diffusion de l'activation ne s'effectue que dans un seul sens : de l'entrée acoustique vers les représentations. Par conséquent, ce modèle accorde une importance particulière aux informations issues du signal. Cependant ce dernier prévoit tout de même que d'autres informations telles que le contexte (e.g., de la phrase) puissent influencer le processus de reconnaissance de mots en intervenant après la formation de la cohorte initiale. Ainsi la version I du modèle de la cohorte présente l'avantage de fournir des hypothèses fortes relatives au moment auquel un mot est censé être reconnu. Ce moment est explicité par le terme de point de reconnaissance ou *point d'unicité* (cf. section 2.2.5 de ce chapitre). Remarquons que bien que l'opérationnalisation de ce point d'unicité soit relativement arbitraire (Kandel & Boë, 1996) et qu'il soit confondu avec la fin acoustique pour une partie des mots de la langue[25], cette mesure semble être corrélée avec les temps de réponse des participants à différentes tâches (e.g., Frauenfelder et al., 1990 , cf. section 2.2.5) (voir aussi Radeau, Morais, & Dewier, 1989).

Cependant, un autre aspect de son architecture a donné naissance à plusieurs critiques. Premièrement, ce dernier postule que pour qu'un candidat soit inclus dans la cohorte initiale, celui-ci doit comporter *exactement* les mêmes premiers phonèmes que le signal d'entrée, un défaut d'appariement à ce niveau l'excluant immédiatement de la cohorte. Or, Connine et al. (1993) ont montré qu'un défaut d'appariement de un ou deux traits

[25] En effet, particulièrement pour les mots courts, le point d'unicité est situé à la fin du mot.

phonétiques pour le phonème initial n'empêchait pas sa reconnaissance (e.g., « panner » pour le mot anglais « banner », bannière, cf. section 2.2.4). Deuxièmement, aucun mécanisme dans l'architecture de la version I du modèle de la cohorte ne permet d'expliquer les données expérimentales montrant qu'un mot fréquent est plus rapidement reconnu qu'un mot rare dans le langage oral (Taft & Hambly, 1986). La version II du modèle de la cohorte présentée ci-dessous tente de résoudre ces différents problèmes.

2.3.1.2. Le modèle de la Cohorte II (Marslen-Wilson, 1987, 1990)

Premièrement, afin de rendre compte de ces critiques, le modèle de la Cohorte II propose, à l'inverse de la version I, un mécanisme d'appariement non plus binaire, en « tout ou rien », mais plutôt graduel, relatif à la *qualité d'ajustement.* Pour cela, il postule que différents *niveaux d'activation* sont attribués à différentes représentations lexicales et en fonction de leur niveau d'appariement avec l'entrée sensorielle. Plus le stimulus a de *traits phonétiques communs* avec une représentation et plus celle-ci reçoit d'activation. Un mot est reconnu lorsque la quantité d'activation dépasse un certain seuil. Cela permet d'expliquer qu'un pseudo-mot tel que « panner » est reconnu comme « banner », en dépit d'un défaut d'appariement relatif à un trait phonétique (ici, le voisement) du phonème initial. Remarquons que dans cette nouvelle version du modèle, ce n'est donc plus le phonème mais bien le trait phonétique qui sert de représentation intermédiaire pour contacter le lexique. Le phonème est considéré comme une unité pouvant émerger *après* l'accès au lexique, c'est-à-dire après qu'un mot soit reconnu. Ainsi, le modèle de la Cohorte II prédit qu'un phonème est plus rapidement reconnu dans un mot plutôt que dans un pseudo-mot uniquement que lorsque le phonème à détecter est situé après le point d'unicité du mot (e.g., Frauenfelder, et al., 1990). En d'autres termes ce modèle prédit une influence du niveau lexical sur le processus de détection des phonèmes uniquement lorsque le mot est déjà reconnu (i.e., après son point d'unicité).

Deuxièmement, la version II de ce modèle permet également de rendre compte des effets de fréquence observés dans la littérature. En effet, en postulant l'existence d'un seuil d'activation fixe d'une part et d' un niveau d'activation à état de repos du système plus élevé pour les mots fréquents que pour les mots rares d'autre part, il permet d'expliquer que les premiers nécessitent moins d'activation (et donc moins de temps) pour être reconnus que les seconds. Notons néanmoins qu'à l'instar de la version I, le modèle de la Cohorte II postule uniquement de l'activation bottom-up et accorde toujours un statut privilégié au début de

mot. Remarquons enfin qu'une implémentation connexionniste de ce modèle cognitif existe (e.g., « Distributed Cohort Model », Gaskell & Marslen-Wilson, 1997).

2.3.2. Le modèle d'activation interactive, TRACE (McClelland & Elman, 1986)

Développé par McClelland & Elman en 1986, TRACE est inspiré d'un modèle à activation interactive décrivant la reconnaissance de mots écrits (McClelland & Rumelhart, 1981). A la différence du modèle de la Cohorte, il postule l'existence de trois étapes lors de l'accès au lexique (cf. Figure 18). Ces étapes correspondent à des niveaux de traitement successifs de l'information auditive. Ainsi, l'ensemble du système est appelé « TRACE » car le pattern d'activation créé par le stimulus sensoriel est une *trace* de l'analyse de cette stimulation d'entrée à chaque niveau de traitement. Chaque niveau est constitué par un ensemble d'unités, correspondant en réalité à un type de détecteurs (i.e., de traits phonétiques, de phonèmes et de mots). Ces unités sont organisées hiérarchiquement avec des connexions *facilitatrices* (ou activation inter-niveaux) situées entre les unités des niveaux inférieur et supérieur (traits-phonèmes, phonèmes-mots).

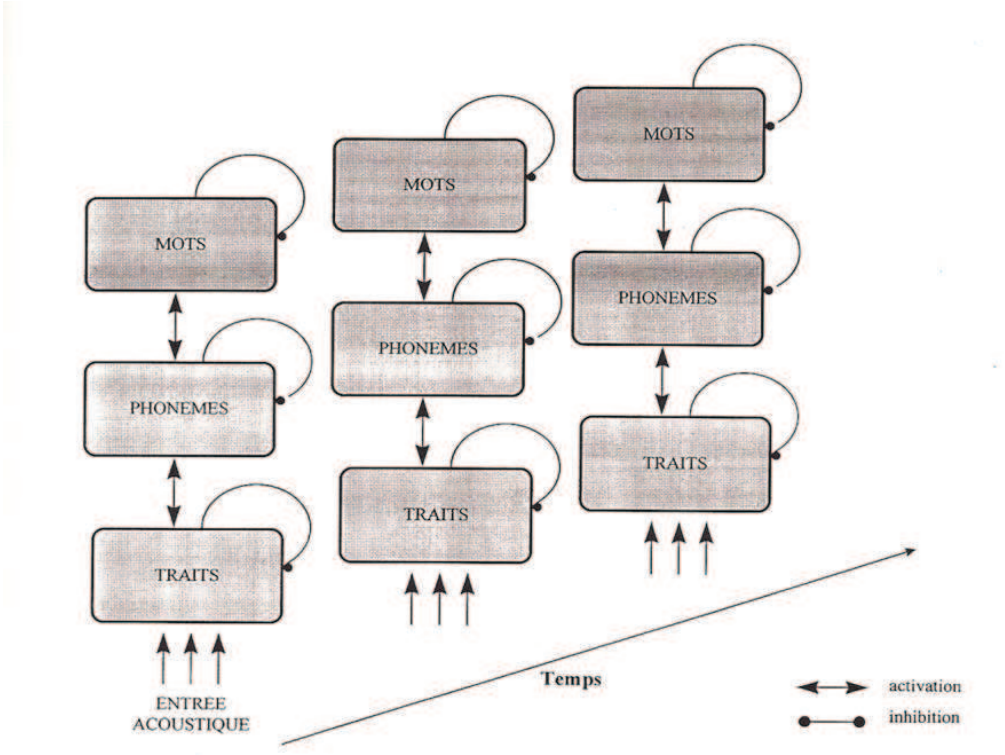


Figure 18. *Représentation schématique du modèle TRACE (McClelland & Elman, 1986). (Extrait de Frauenfelder, 1996).*

A la différence du modèle de la Cohorte II, TRACE postule que ces connexions *facilitatrices* inter-niveaux diffusent de l'activation de manière *bilatérale* : de manière *ascendante*, des traits vers les phonèmes puis des phonèmes vers les mots mais aussi de manière *descendante*, des mots vers les phonèmes et des phonèmes vers les traits (cf. doubles flèches de la Figure 18). Ce réseau est hautement interconnecté puisqu'il dispose également de connexions *inhibitrices* entre les unités d'un même niveau de traitement (ou inhibition intra-niveau).

Selon ce modèle, lorsqu'un signal acoustique de parole arrive à notre système perceptif, il est tout d'abord analysé en traits phonétiques (place d'articulation, mode d'articulation, etc.) par les détecteurs (ou unités) correspondants. C'est au second niveau que l'ensemble des traits phonétiques activés lors de l'étape précédente (e.g., place d'articulation bilabiale) permettent d'activer l'ensemble des unités phonémiques partageant des caractéristiques phonétiques (e.g., l'ensemble des phonèmes bilabiaux). De plus, l'activation d'un phonème active tous les mots au niveau lexical qui contiennent ce phonème et ce avec la même intensité *quel que soit sa position* dans le mot. La quantité d'activation avec laquelle un ensemble d'unités (e.g., traits) active une unité à un niveau supérieur (e.g., un phonème) dépend du niveau d'activation des unités de départ (e.g., traits). Ce degré d'activation des unités de départ détermine également la force de connexion entre les unités d'où l'activation est issue (e.g., traits) et l'unité correspondante au niveau de traitement suivant (e.g., un phonème). Il existe donc une relation proportionnelle entre la force des connexions inter-niveaux et la quantité d'activation fournie par l'entrée sensorielle.

Comme nous l'avons évoqué plus haut, le système postule que l'activation diffuse de manière ascendante mais également descendante entre les niveaux. Ainsi, lorsqu'un ensemble d'unités (e.g., mots) est activé, celui-ci renvoie ensuite de l'activation sur les unités du niveau inférieur (e.g., phonèmes) le composant. Ce mécanisme permet de renforcer les unités déjà activées. Il permet également d'expliquer les effets de supériorité du mot observés dans plusieurs études expérimentales (e.g., Frauenfelder, et al., 1990; Ganong, 1980; Samuel, 1981). En effet TRACE postule que lorsqu'un phonème est situé dans un mot, l'unité phonémique correspondant reçoit de l'activation ascendante mais également de l'activation descendante. En revanche, lorsqu'un phonème est situé dans un pseudo-mot, elle ne peut recevoir que de l'activation ascendante, les pseudo-mots ne disposant pas de représentations au niveau lexical. Ainsi, ce mécanisme d'activation descendante permet d'expliquer le fait qu'un phonème est plus rapidement reconnu dans un mot plutôt que dans un pseudo-mot, ce dernier recevant plus d'activation dans le premier cas que dans le second.

De surcroît, en parallèle de ces mécanismes d'activation, de l'inhibition circule également entre les unités d'un même niveau. Par exemple lorsqu'un phonème est activé, tous les autres phonèmes sont inhibés, ces derniers étant situés au même niveau. Il en est de même pour chaque étape de traitement. Ce degré d'inhibition est, à l'instar du degré d'activation, proportionnel à la force des connexions et du niveau d'activation initial de l'unité à l'origine de l'inhibition. Plus une unité est activée par une (des) connexion(s) inter-niveaux et plus celle-ci inhibe fortement ses voisines au même niveau. C'est par l'action conjointe des mécanismes d'activation descendante et d'inhibition que des unités sont peu à peu plus activées que les autres. Ainsi, un mot est reconnu lorsqu'une unité lexicale a un niveau global d'activation significativement[26] plus élevé que toutes ses voisines. Par conséquent, alors que le modèle de la Cohorte II explique l'affinement des hypothèses lexicales (ou compétition lexicale) en termes de *réduction* du nombre de candidats, TRACE postule plutôt qu'un mécanisme de *différenciation* serait à l'origine de ce phénomène. Notons que comme le modèle de la Cohorte II, TRACE permet de rendre du compte des effets de fréquence en définissant un seuil d'activation de base plus bas pour les mots fréquents que pour les mots rares.

Globalement, nous pouvons remarquer que la notion de degré et de seuil d'activation est centrale dans le modèle TRACE. C'est également en postulant une diffusion d'activation bilatérale et des connexions inhibitrices intra-niveaux que ce modèle permet de tolérer des défauts d'appariement entre le stimulus d'entrée et sa représentation lexicale tout en conservant le *phonème* comme unité fonctionnelle permettant d'accéder au lexique. Notons que contrairement au modèle de la Cohorte II, TRACE n'accorde pas de statut privilégié au début de mot. Notons également que le postulat d'une boucle rétroactive (i.e., activation descendante) permettant d'expliquer les effets lexicaux constitue l'une des plus grandes singularités de ce modèle.

2.3.3. Le modèle Shortlist

2.3.3.1. Shortlist A (Norris, 1994)

Le modèle Shortlist A est un modèle connexionniste qui a été élaboré afin de remédier à certaines imperfections de TRACE. La différence majeure entre son architecture et celle de TRACE est que celui-ci ne postule l'existence d'aucune boucle de rétroaction pour

[26] Cette différence est considérée comme significative lorsque celle-ci dépasse un certain seuil, déterminé par la règle de Luce (Luce, 1959), cité dans Spinelli (1999).

expliquer l'influence du lexique sur le traitement de la parole. Dans ce modèle, la totalité de l'activation diffuse uniquement de manière ascendante entre les différents niveaux de décodage, c'est-à-dire du signal vers les unités pré-lexicales puis des unités pré-lexicales vers les représentations de mots. La seconde différence avec le modèle TRACE est que Shortlist A postule deux étapes temporellement distinctes dans le processus d'accès au lexique. Pour TRACE, le processus d'activation des candidats lexicaux ainsi que celui de la compétition entre les unités s'effectue de manière simultannée, ce qui oblige ce dernier à supposer un mécanisme coûteux de duplication du lexique à chaque nouveau phonème perçu.

Pour Shortlist A, il existe donc deux étapes successives pour l'accès au lexique. La première correspond à une phase de recherche exhaustive de candidats lexicaux compatibles avec l'entrée sensorielle. Lors de cette phase, un nombre limité de représentations lexicales (ou « short-list[27] ») qui partagent toutes plusieurs caractéristiques phonémiques avec le signal acoustique est activé. Cette « short-list » est établie sur la base d'un score d'ajustememement, ce qui permet de rendre compte de la compatibilité du stimulus d'entrée avec les différentes représentations. Ce score d'appariement est effectué sur la base du nombre de *phonèmes* partagés entre le signal d'entrée et chaque unité lexicale. Cette notation est effectuée en atribuant un nombre positif (i.e., + 1) pour le partage de chaque phonème ; dans le cas contraire, une cotation négative (i.e., -3) est associée. L'exemple donné par les auteurs pour illustrer cette attribution d'un score est le suivant : lorsque le signal d'entrée correspond à /k/ /æ/ /t/, les mots /kæt/ « cat », chat et /kætəlɒg/ « catalog », catalogue obtiennent le même score (+1 +1 +1) = 3 ; les mots /kæp/ « cap », casquette et /kæptən/ « captain », capitaine sont quant à eux pénalisés de la même manière pour le défaut d'appariement au niveau du troisième phonème (+1 +1 -3) = -1. Notons ici qu'à la différence du modèle de la Cohorte II, Shortlist A ne pénalise pas plus fortement un défaut d'appariement pour le premier phonème du stimulus que pour un phonème ayant une autre position dans le mot : il n'accorde donc pas de statut privilégié au début de mot lors de sa reconnaissance. Durant cette seconde étape, la « short-list » est importée dans un réseau de compétition lexicale (« lexical competition network »). La principale caractéristique de cette phase est que les unités lexicales qui vont entrer en compétition sont celles ayant reçues, lors de la première phase, de l'activation pour la même section de l'entrée sensorielle. En d'autres termes, cela signifie que seules les représentations disposant d'au moins un phonème en commun vont entrer en compétition. Pour rendre compte de ce processus, Shortlist A postule, à l'instar de TRACE, des connexions inhibitrices directes entre ces unités (inhibition

[27] Littéralement : « courte liste »

intra-niveau). La Figure 19 représente un exemple de ces connexions pour la reconnaissance du mot « catalog ».

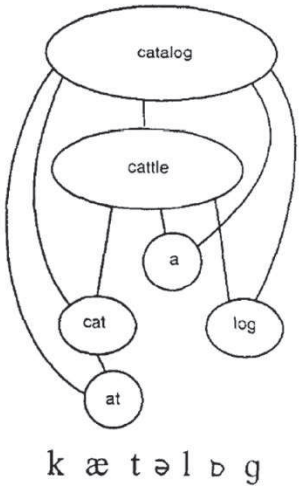


Figure 19. *Exemple de connexions inhibitrices entre les candidats lexicaux générés lors de la production du mot « catalog ». Remarquons que cette figure ne montre pas la totalité des candidats pouvant entrer en jeu dans ce processus. Le cas échéant, les mots « battle », « catalyst », etc. devraient également être inclus. (Extrait de Norris, 1994).*

Globalement, nous pouvons remarquer dans Shortlist A que, l'attribution de scores d'appariement permet de rendre compte de la reconnaissance d'un mot, même si celui-ci est mal prononcé (e.g., « chigarette »). Ce modèle permet ainsi de tolérer des défauts d'appariement entre le stimulus d'entrée et sa représentation lexicale tout en conservant le *phonème* comme unité fonctionnelle permettant d'accéder au lexique. Ensuite contrairement au modèle de la Cohorte II, Shortlist A n'accorde pas de statut privilégié au début de mot. Enfin, Shortlist A postule de l'activation et de l'inhibition ascendante, mais pas, à la différence de TRACE, d'activation ou d'inhibition descendante. Les résultats indiquant une influence du niveau lexical sur le processus de détection de phonèmes sont expliqués par le fait que le modèle suppose comme que deux « routes » peuvent être empruntées pour décoder un signal de parole : une « route phonémique » allant vers les unités phonémiques pré-lexicales ainsi qu'une route lexicale, permettant de contacter les représentations de mots contenues dans le lexique. Ce concept de double voie a initialement été développé par le modèle Race (Cutler & Norris, 1979). Afin qu'un phonème puisse être identifié à partir de la route lexicale comme de la route phonémique, Shortlist A postule explicitement que ces mots disposent également de représentations phonologiques. Ce modèle permet d'expliquer qu'un phonème est plus rapidement reconnu dans un mot (empruntant la voie lexicale) que dans un pseudo-mot (empruntant la voie phonémique), la voie lexicale permettant de produire des réponses concernant l'identité d'un phonème plus rapidement que la voie phonémique.

2.3.3.2. Shortlist B (Norris & McQueen, 2008)

Le modèle Shorlist B a été récemment proposé par Norris et McQueen en 2008. Il est basé sur le modèle Shortlist A et partage de nombreux postulats avec son prédécesseur. En effet, celui-ci postule toujours deux étapes distinctes pour l'accès au lexique : une phase de recherche de candidats lexicaux ainsi qu'une phase d'évaluation compétitive. De plus, il suppose que le processus d'accès au lexique s'effectue uniquement grâce à un processus ascendant, de l'entrée vers les représentations de mots. Cependant, Shortlist B se distingue de Shortlist A sur une caractéristique principale. En effet, tandis que Shortlist A est un modèle connexionniste basé sur de la diffusion d'activation, Shortlist B repose sur des calculs *probabilistes* Bayésiens[28]. Lors du processus d'accès au lexique le modèle calcule la probabilité conditionnelle pour chaque représentation lexicale compte tenu de l'information disponible dans le signal, selon la formule suivante :

$$p\ Word_i\ Evidence\ = \frac{p\ Evidence\ Word_i\ * p(Word_i)}{\sum_{j=1}^{n} p\ Evidence\ Word_j\ * p(Word_j)}$$

Avec :

$p(Word_i)$ = probabilité associée au mot i

$Evidence$ = signal d'entrée

$p(Word_j)$ = probabilité associée au mot j

$p\ Evidence\ Word_i\ =$ probabilité pour le signal d'entrée en sachant que mot i est reconnu

$p\ Evidence\ Word_j\ =$ probabilité pour le signal d'entrée en sachant que mot j est reconnu

Un mot est reconnu lorsque la probabilité qui lui est associée excède un certain seuil. Ainsi, le modèle Shortlist B permet de rendre compte des mêmes effets que son prédécesseur. De plus, il permet également de simuler les effets de fréquences en estimant la probabilité a priori associée à un mot (i.e., $p(Word_i)$) à partir de la fréquence d'occurrence répertoriée dans les bases de données. Notons également que Shortlist B permet aussi d'expliquer les effets de densité de voisinage phonologique sur le processus de reconnaissance de mots (voir Luce & Pisoni, 1998, pour une revue) indiquant qu'un mot avec peu de voisins et/ou des voisins plus rares que lui sera plus rapidement reconnu qu'un mot avec une densité de voisinage phonologique importante et dont les voisins sont plus fréquents. Remarquons que par conséquent, à la différence de Shortlist A, l'absence de toute activation/inhibition dans le

[28] Les probabilités Bayésiennes sont issues du théorème de Bayes, la première fois décrit par Thomas Bayes. Ce théorème permet de prédire la probabilité conditionnelle d'un évènement A sachant que l'évènement B s'est déjà produit.

modèle va engendrer un processus de compétition non plus direct (i.e., effectué par des connexions inhibitrices entre les représentations) mais indirect résultant de calculs probabilistes pour chaque représentation de mot.

2.3.4. Le modèle Merge (Norris, et al., 2000)

Le modèle Merge est une version de Shortlist A qui a été spécifiquement développée pour rendre compte des effets de supériorité du mot obsevés lors de la tâche de détection de phonèmes. Ce dernier a été proposé par Norris et al. (2000). La différence majeure entre son architecture et celle de TRACE est que celui-ci ne postule l'existence d'aucune boucle de rétroaction pour expliquer les effets lexicaux évoqués dans la section 2.2.5. En effet, Norris et collègues postulent qu'aucune activation descendante n'est nécessaire lorsqu'il s'agit notamment d'expliquer l'effet de supériorité du mot obtenu en détection de phonèmes (e.g., Frauenfelder, et al., 1990; Rubin, et al., 1976) : « Feedback is never necessary » (Norris et al., 2000, p. 299). Pour cela, les auteurs proposent le modèle Merge, décrivant l'accès au lexique en ne postulant que des connexions *ascendantes* (cf. Figure 20). Cette conception, dite « autonome » du processus de reconnaissance de mots est également issue du concept de double voie développé dans le cadre du modèle Race (Cutler & Norris, 1979). C'est dans cet esprit que le modèle Merge distingue deux types d'informations : lexicales et pré-lexicales.

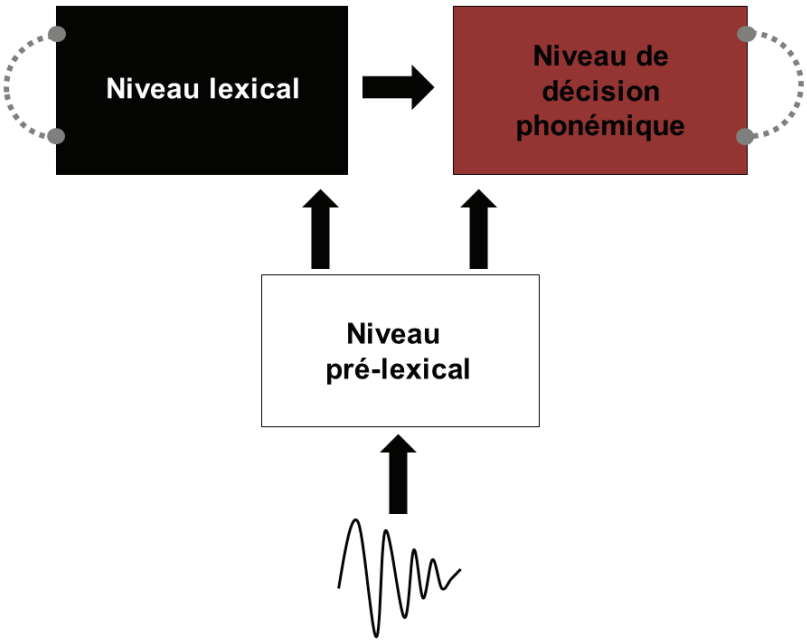


Figure 20. *Représentation schématique de l'architecture générale et des différents modules du modèle Merge (Norris, et al., 2000). Le rectangle du bas représente le niveau des traits pré-lexical (« input node »), celui en haut à gauche correspond au niveau lexical (« lexical node ») alors que celui en haut à droite représente le niveau de décision phonémique (« phoneme decision node »). Les connections inter-niveaux sont excitatrices et unidirectionnelles (flèches pleines), les connections intra-niveaux n'opèrent qu'aux niveaux lexical et de décision phonémique et sont inhibitrices et bidirectionnelles (lignes pointillées). (Adapté et simplifié de Norris, et al., 2000).*

Dans ce modèle, trois modules (ou nœuds) sont représentés, chacun composé par un ensemble d'unités : l' « input node » (correspondant au niveau pré-lexical), le « lexical node » (correspondant au niveau lexical) ainsi que le « phonemic decision node » (correspondant au niveau de décision phonémique). Ainsi, lorsqu'un stimulus de parole est perçu, celui-ci active d'abord le niveau d'entrée du signal acoustique. Ensuite, cette activation diffuse à la fois vers le niveau lexical (voie lexicale) et vers le niveau de la décision phonémique (voie phonémique). Enfin, l'activation issue du niveau lexical se propage vers le niveau de décision phonémique. C'est grâce à ce mécanisme que Merge permet de rendre compte du fait qu'un phonème est plus rapidement reconnu dans un mot que dans un pseudo-mot. En effet, les unités de décision phonémique reçoit plus d'activation dans le premier cas que le second. Or, contrairement à TRACE, le niveau lexical n'est pas situé à un niveau « supérieur » de traitement par rapport aux unités de décisions phonémiques. Au contraire, c'est le module responsable du traitement des phonèmes qui intervient en tant que mécanisme décisionel, *après* l'accès au lexique. La diffusion d'activation du niveau lexical vers le niveau de décision phonémique est donc uniquement ascendante. Cela permet à Merge de rendre compte des effets lexicaux observés dans la littérature, sans postuler le moindre mécanisme retroactif.

A ce propos, signalons ici que Merge ne suppose l'existence de connexions effectives entre le niveau de décision phonémique et les autres modules que lorsque la tâche du participant exige une opération *explicitement* relatives aux phonèmes (e.g., tâche de détection de phonèmes). Dans le cas contraire, aucune connexion n'est établie avec ce niveau et l'activation diffuse uniquement par la voie lexicale.

Merge postule également des connexions inhibitrices intra-niveaux, présentes au niveau lexical et au niveau de la décision, qui permettent d'inhiber les unités les moins activées et donc de sélectionner celles qui correspondent le mieux au signal d'entrée. De la même manière que dans TRACE, un mécanisme de *différenciation* serait donc à l'origine du phénomène de compétition lexicale.

En résumé, à la différence de TRACE mais de la même manière que le modèle de la Cohorte II, Merge postule un accès direct aux unités lexicales sans passer par le niveau de décision phonémique ainsi qu'un sens de diffusion de l'activation strictement ascendant. Merge étant un modèle connexionniste disposant d'une phase d'apprentissage, ce dernier permet de rendre compte des effets de fréquence en postulant des poids de connexion plus importants du niveau des traits vers les unités lexicales plus les mots sont fréquents. Notons enfin que contrairement au modèle de la Cohorte II, Merge n'accorde pas de statut privilégié au début de mot.

2.3.5. Le modèle NAM : "Neighborhood Activation Model" (Luce & Pisoni, 1998)

Une des caractéristiques du modèle NAM est qu'il ne postule pas, à l'instar de TRACE ou de Merge et à la différence du modèle de la Cohorte, une importance spécifique au début de mot lors de la phase d'activation des multiples candidats lexicaux. En effet, dans ce modèle, les représentations compatibles avec au moins une portion de l'entrée auditive vont être activées par l'arrivée d'un signal acoustique de parole et ce quel que soit la position de cette portion de signal en commun dans le stimulus d'entrée. Cependant, la spécificité de NAM est qu'il effectue une prédiction forte quant au type de candidats activés lors de cette phase : il postule que seules les représentations lexicales (ou *voisins phonologiques*) partageant les *mêmes phonèmes à n/+ ou -1 près* avec l'entrée sensorielle seront activées (voir Luce & Pisoni, 1998, pour une revue en faveur de cette hypothèse). Ainsi, le nom « Neighborhood Activation Model » qui signifie littéralement « modèle à activation du voisinage » vient de cette hypothèse. La densité du voisinage phonologique correspond au nombre de voisins phonologiques d'un mot. Elle se calcule en regroupant le nombre de mots pouvant être dérivés du mot original (e.g., /bal/, « bal ») par addition (e.g., /balɛ/, « balai »), substitution (e.g., /kal/, « cale ») ou délétion d'un phonème (e.g., /al/, « halle »). Cependant, à la différence des autres modèles, il ne suppose pas que le signal acoustique correspondant au mot « bal » puisse activer le mot /baldakɛ̃/, « baldaquin ». En effet, ce modèle est à la base conçu pour rendre compte de la reconnaissance de mots monosyllabiques et donc relativement courts.

De la même manière que le modèle de la Cohorte et Merge, NAM postule que l'activation ne va diffuser que de manière ascendante, de l'entré sensorielle vers les unités de décision du mot. Pour NAM, lors de la présentation d'un signal de parole, celui-ci va d'abord activer un certain nombre de caractéristiques acoutico-phonétiques en mémoire. A cette étape, le modèle postule explicitement que les caractéristiques contenues dans le stimulus d'entrée sont activées, que celles-ci soient présentées dans un mot, un pseudo-mot, ou un non-mot. Il s'agit du niveau d'analyse pré-lexical du traitement de la parole.

Ensuite, l'activation issue de ce niveau va diffuser vers les unités de décision du mot. Ainsi, seules les unités de décision correspondant à un *mot* en mémoire vont être activées. La particularité de ce modèle est qu'il postule que la présentation d'un mot-cible va venir activer sa représentation mais également toutes celles correspondant à des mots phonétiquement proches de lui, ses voisins phonologiques. La quantité d'activation reçue par les voisins phonologiques étant identique à celle du mot-cible à ce niveau, les voisins vont dans un

second temps entrer en *compétition* avec la représentation correspondant au mot-cible. Cette étape de compétition n'est pas formalisée par des mécanismes d'inhibitions mais plutôt par des calculs probabilistes effectués au niveau de *décision* du mot permettant d'estimer différentes *valeurs décisionnelles* pour chacune des candidats activés. Ces calculs sont effectués en tenant compte de quatre paramètres : la probabilité associée au mot, sa fréquence d'occurrence, la probabilité ainsi que la fréquence associée à chacun de ses voisins. Notons que la probabilité associée au mot et à ses voisins correspond à la quantité d'activation reçue au niveau des caractéristiques acoustico-phonétiques leur correspondant. Ainsi, la valeur décisionnelle pour chaque unité de mot est exprimée par le rapport suivant :

$$p\ ID\ = \frac{SWP * Freq_s}{SWP * Freq_s + \ \sum_{j=1}^{n}[NWP_j * Freq_j]}$$

Avec :

SWP = probabilité associée au mot stimulus

$Freq_s$ = la fréquence du mot

NWP_j = la probabilité associée au voisin j

$Freq_j$ = la fréquence du voisin j

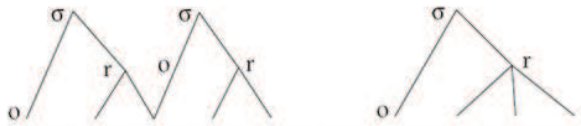
De ce fait, la probabilité d'identifier un mot va dépendre de sa fréquence et de la probabilité qui lui est associée mais également, du nombre de ses voisins et de leur probabilité associée, ainsi que de la fréquence de ces voisins. Ainsi, plus un mot est fréquent, moins il a de voisins et moins ces voisins sont fréquents par rapport à lui-même et plus ce dernier va être facilement identifié. Un mot va être reconnu lorsque la valeur décisionnelle qui lui est attribuée va être suffisamment importante pour dépasser un certain critère, lui permettant de passer alors en mémoire de travail.

En résumé, ce qui rend NAM singulier c'est qu'à la différence des autres modèles cités ci-dessus, il postule que la densité du voisinage phonologique est un paramètre influençant *directement* le processus de reconnaissance de mots. De ce fait, il n'accorde pas d'importance particulière au début de mot. Notons cependant, qu'il considère que la fréquence d'occurrence des mots ne va pas, contrairement à la densité du voisinage phonologique, influencer directement l'étape d'activation initiale des candidats à partir du signal d'entrée. En effet, il postule que cette influence va intervenir plus tard dans le processus, à un niveau post-lexical, lors du calcul des valeurs décisionnelles au niveau des

unités de décision du mot. Il permet donc d'expliquer les effets de fréquence retrouvés dans la littérature par un mécanisme *post-lexical*, bien que celui-ci intervienne avant que la reconnaissance du mot ne soit achevée. Ensuite, NAM suppose que ce sont les caractéristiques acoustico-phonétiques de base qui sont à l'origine du contact avec les représentations de mot. Enfin, à l'instar du modèle de la Cohorte et de Merge, il fait l'hypothèse que des mécanismes strictement ascendants sont à l'origine du processus de reconnaissance de mots. Toutefois, remarquons que malgré le fait qu'un grand nombre de preuves indiquent que le processus de reconnaissance de mots semble influencé par la densité du voisinage phonologique (voir Luce & Pisoni, 1998, pour une revue), ce modèle dispose d'une importante limitation. En effet, ce dernier a uniquement été échafaudé sur la base d'études perceptives où les participants avaient pour tâche d'identifier des mots dans du bruit. De ce fait, NAM est basé sur des mesures effectuées en temps différé et ne prend pas en compte le décours temporel du traitement de l'information lexicale dans son architecture, alors qu'il s'agit d'un paramètre important de l'accès au lexique (cf. section 2.2.5 ; voir Norris & McQueen, 2008, pour une critique similaire). Remarquons enfin qu'une implémentation connexionniste de NAM existe (e.g., « Parsyn», Luce, Goldinger, Auer, & Vitevitch, 2000).

2.3.6. Le modèle LAFF : "Lexical Access From Features" (Stevens, 2002)

Le modèle LAFF est issu de la « théorie quantique de la perception de la parole », proposé par Stevens et collègues (Stevens, 1989; Stevens & Keyser, 2010). Comme son nom l'indique, un des postulats majeurs de ce modèle est qu'il suppose, à l'instar de Merge et du modèle de la Cohorte II, un accès aux unités lexicales par l'intermédiaire des *features* ou caractéristiques/traits acoustico-phonétiques et de la syllabe, sans médiation par le phonème en tant qu'unité autonome (Stevens, 2005). Soulignons ici que comme pour Merge, l'identification consciente du phonème n'est possible qu'à un niveau de traitement décisionnel, après que le mot entier soit reconnu. Ainsi, chaque représentation de mot dans le lexique va être définie par un ensemble de *features* et par sa structure syllabique et non par un ensemble de phonèmes (cf. Figure 21).



	s	ʌ	d	ə	n		h	ɛ	l	p
vowel		+		+				+		
glide							+			
consonant	+		+		+				+	+
stressed		+		−				+		
reducible		−		+				−		
continuant	+		−		−				−	−
sonorant			−		+				+	−
strident	+									
lips										+
tongue blade	+		+		+				+	
tongue body										
round										−
anterior	+		+		+				+	
lateral									+	
high		−		−				−		
low		−		−				−		
back		+		+				−		
tense		−						−		
spread glottis							+			
nasal					+					
stiff vocal folds	+		−							+

Figure 21. *Représentations lexicales des mots anglais « sudden », (soudain) et « help », (aide) postulées par le modèle LAFF. La colonne la plus à gauche désigne les différents types de « features ». La structure syllabique de chaque mot est schématisée au sommet du tableau : « σ » correspond à la totalité, « o » à l' « onset » ou au début et « r » à la rime ou la fin de la syllabe. (Extrait de Stevens, 2005).*

Ce modèle est issu du domaine de la phonétique et met de ce fait l'accent sur les étapes de décodage de ces traits ainsi que sur leur classifiation. Ainsi, à la différence des des modèles explicités ci-dessus, LAFF postule explicitement *plusieurs* étapes permettant d'associer le *signal* acoustique à différentes *features* correspondant (cf. Figure 22). Néanmoins, de la même manière que les autres modèles présentés ci-dessus, LAFF décrit qu'un mot va être reconnu en postulant un appariement entre le stimulus d'entrée et les représentations de mots que l'on a en mémoire. La qualité de cet appariement va déterminer le niveau d'activation de chacune de ces représentations en mémoire.

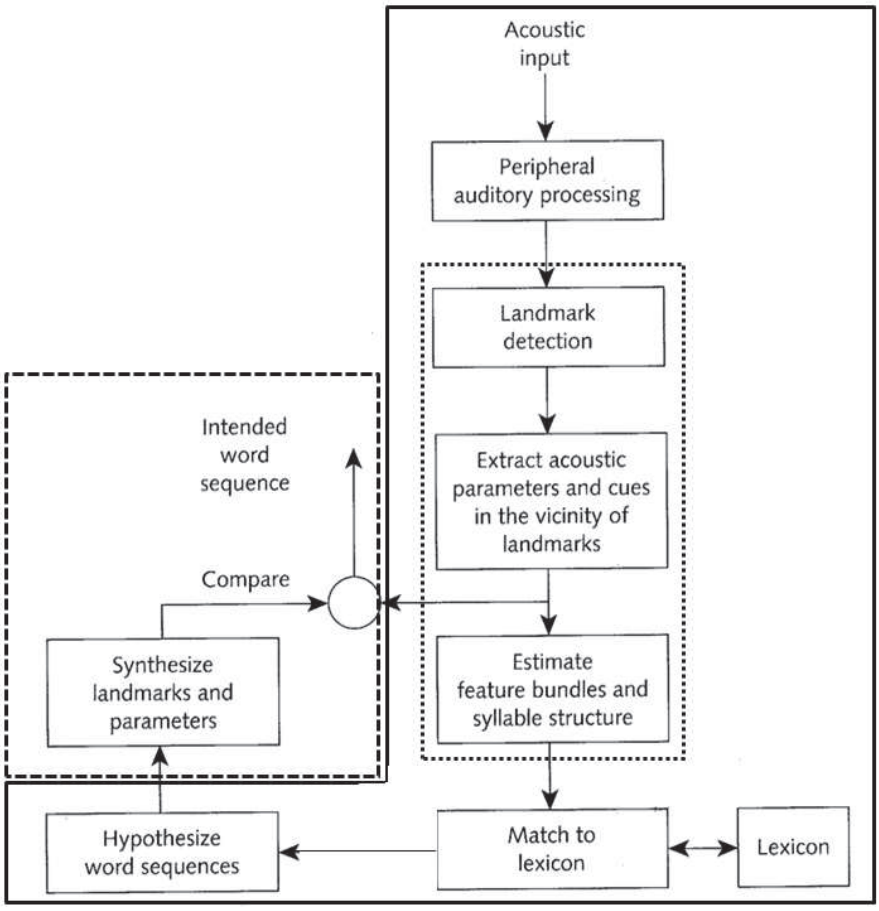


Figure 22. *Représentation schématique des étapes du traitement du signal acoustique permettant l'accès au lexique selon le modèle LAFF. (Extrait de Stevens, 2005).*

LAFF décrit deux étapes distinctes pour l'accès au lexique. La première (encadrée par la ligne continue sur la Figure 22) est une phase ascendante où l'activation diffuse du signal vers le lexique. Lors de cette étape, des « landmarks » ou « repères » sont premièrement détectés dans le signal acoustique (« Landmark detection »), permettant de distinguer la présence de segments vocaliques, semi-vocaliques et consonantiques. Ces repères correspondent à des patterns de changements d'amplitude dans différentes bandes de fréquences. Ensuite, dans chacune des « zones » délimitées par ces repères, des indices acoustiques plus fins sont extraits (« Extract acoustic parameters and cues in the vinicity of landmarks »), permettant d'inférer : la forme du conduit vocal au-dessus (e.g., place d'articulation d'une consonne) et au niveau du larynx (e.g., absence ou présence de vibration des cordes vocales, fréquence fondamentale). Ce sont ces paramètres acoustiques qui par la suite permettent d'activer différents groupes de *features* et différents types de structures syllabiques (« Estimate feature bundles and syllable structure ») pouvant correspondre au signal d'entrée. C'est sur la base de ces premières estimations qu'un certain nombre de candidats lexicaux sont générés (i.e., « hypothesize word sequences »). La seconde étape

(encadrée par les grands pointillés sur la Figure 22) correspond à une étape d' « analyse par synthèse » ou l'activation diffuse de manière descendante (i.e., du lexique vers les unités pré-lexicales) et permet de comparer les *features* des candidats lexicaux activés avec les paramètres acoustiques du signal d'entrée. Cette opération a pour objectif de réduire le nombre d'hypothèses lexicales, jusqu'à ce qu'un mot soit reconnu. Notons que c'est uniquement lors de cette étape que des informations d'ordre contextuel (phrase, etc.) peuvent influencer le processus de reconnaissance de mot.

Remarquons que ce modèle se concentre tout particulièrement sur les phases de décodage phonétique du signal de parole (Stevens, 2002), précédant l'accès au lexique (encadrée par les petits pointillés, cf. Figure 22) et non sur le processus d'activation des représentations lexicales. Ainsi, l'importance du début de mot ainsi que l'existence de connections (inhibitrices ou excitatrices) entre les différentes unités d'un même niveau ne sont pas clairement spécifiées. Du fait de ce manque d'information, nous n'utiliserons pas ce modèle afin d'interpréter les résultats de nos propores études (cf. Chapitre 3, 4 et 5). Nous avons néanmoins décidé de présenter ce modèle dans cette section car celui-ci évoque, à la différence des autres décrits ici, la possibilité d'inclure l'information visuelle (relative aux gestes articulatoires du locuteur) dans son architecture. En effet, LAFF postule que chaque « feature » est associée à une représentation acoustique mais également articulatoire. Le rôle donné à l'information visuelle resterait néanmoins subsidiaire et viendrait particulièrement améliorer la recherche des mots dans le lexique lorsque le signal acoustique est détérioré « other information may be available to the listener in addition to that derived from analysis of the acoustic signal. This information includes visual cues derived from observation of the speaker's face [...]. Cues of this type could greatly aid in the search for words in the lexicon, particularly in the presence of noise » (Stevens, 2005, p. 151).

2.3.7. Conclusions sur les modèles d'accès au lexique

En conclusion, nous avons pu tout d'abord observer que les modèles de reconnaissance de mots parlés présentés dans cette section reposent sur plusieurs postulats communs. En effet, l'ensemble d'entre eux propose l'existence d'unités pré-lexicales permettant d'apparier le signal acoustique avec des représentations abstraites de mots en mémoire. Ces modèles supposent également la présence de phases d'activations multiple puis de compétition/sélection entre différents candidats lexicaux, afin de modéliser le processus de reconnaissance de mots. Cependant, malgré ces ressemblances d'ordre général, nous

avons vu que ces derniers effectuent des prédictions bien différentes (cf. Tableau 1) quant au type d'unité (e.g., phonétiques, phonémiques, etc.) permettant d'accéder au lexique ainsi qu'au sens des connexions impliquées entre les différents stades de traitement de l'information auditive (connexions ascendantes, descendantes, etc.). Egalement, nous avons vu que ces derniers n'accordent pas tous une importance particulière au début de mot dans le processus d'accès au lexique. Ces propriétés seront discutées en fonction des résultats des travaux proposés dans la partie expérimentale de ce manuscrit.

Tableau 1. *Résumé des principales caractéristiques des modèles décrivant l'accès au lexique évoqués dans ce chapitre. La compétition lexicale directe fait référence à la présence de connexions inhibitrices intra-niveau au niveau lexical.*

Modèle	Diffusion d'activation	Unité fonctionnelle	Importance de début de mot	Compétition lexicale directe
Cohorte I	Ascendante	Phonème	Oui	Non
Cohorte II	Ascendante	Trait	Oui	Non
TRACE	Bidirectionnelle	Phonème	Non	Oui
Shortlist A	Ascendante	Phonème	Non	Oui
Shortlist B	Ø (Bayésien)	Phonème	Non	Non
Merge	Ascendante	Niveau pré-lexical	Non	Oui
NAM	Ascendante	Caractéristiques acoustico-phonétiques	Non	Non
LAFF	Bidirectionnelle	Trait *(features)*	Non	Non renseigné

2.4. BASES NEURALES DU TRAITEMENT DE L'INFORMATION LEXICALE

Dans cette section nous allons passer en revue différents travaux qui ont cherché à distinguer l'existence de régions ou de réponses cérébrales spéciquement impliquées dans le traitement de la parole à un niveau lexical par opposition à un niveau pré-lexical. Nous renvoyons le lecteur à d'autres travaux pour une perspective relative au décours temporel du processus de reconnaissance de mots (Pulvermüller, 2007). Nous allons donc présenter plusieures études ayant comparé l'activité cérébrale de participants en train d'effectuer différentes tâches sur du matériel lexical (e.g., mots) et non lexical (e.g., pseudo-mots, non-mots, etc.). Notons que l'ensemble de ces travaux concerne la reconnaissance de mots *parlés* isolés, en modalité auditive seule.

Ainsi, il est tout d'abord possible d'effectuer une distinction parmis ces travaux, opposant certaines études qui m'obtiennent aucune différence entre le traitement des mots et des pseudo-mots (e.g., Binder et al., 2000) à celles qui en observent (e.g., Klein, Moeller, Nuerk, & Willmes, 2010; Kotz, Cappa, Cramon, & Friederici, 2002; Majerus et al., 2002; S. D. Newman & Twieg, 2001; Orfanidou, Marslen-Wilson, & Davis, 2006; Raettig & Kotz, 2008; Xiao et al., 2005). Dans ce dernier groupe, selon Raettig et Kotz (2007), il est possible d'opposer les études ayant trouvé une différence en termes de quantité d'activation pour des régions similaires (plus d'activation pour le décodage des mots : Orfanidou et al., 2006 ; ou plus d'activation pour le traitement des pseudo-mots : Newman et al., 2001) à ceux suggérant une augmentation de l'activation pour des zones différentes pour les mots et les pseudo-mots (Klein, et al., 2010; Kotz, et al., 2002; Majerus, et al., 2002; Raettig & Kotz, 2008; Xiao, et al., 2005). A première vue, ces études semblent obtenir des résultats très différents. Or, cette variabilité peut (au moins en partie) être due à l'utilisation de tâches différentes (e.g., écoute passive, détection de phonèmes, tâche de décision lexicale, etc.), mais également à l'utilisation de stimuli lexicaux (e.g., nombres, Klein, et al., 2010) et non lexicaux (e.g., pseudo-mots et non-mots) impliquant une activité cérébrale différente (e.g., pseudo-mots vs. non-mots, e.g., Specht, 2003). Cependant, lorsque l'on considère uniquement les données indiquant une augmentation de l'activité cérébrale pour les mots, certaines zones cérébrales communes semblent émerger (cf. Figure 23), alors que ce n'est pas le cas lorsque les zones cérébrales indiquant une augmentation de l'activité pour le traitement des pseudo-mots sont considérées (Raettig & Kotz, 2008).

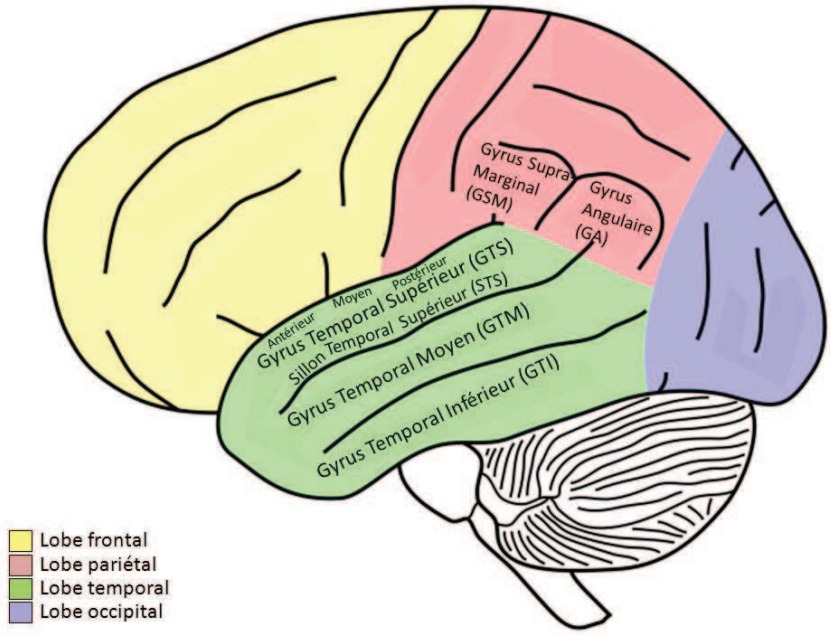


Figure 23. *Représentation schématique des principales régions impliquées dans le décodage de l'information lexicale.*

Ainsi, plusieurs études semblent avoir observé une activation spécifique pour le traitement des stimuli lexicaux dans les zones temporales moyennes (i.e., au niveau du gyrus temporal moyen, GTM : Kotz, et al., 2002; Orfanidou, et al., 2006; Raettig & Kotz, 2008) et inférieures (i.e., au niveau du gyrus temporal inférieur, GTI : Majerus, et al., 2002; Orfanidou, et al., 2006; Raettig & Kotz, 2008), mais également à la jonction du lobe temporal et pariétal (i.e., au niveau du gyrus supra-marginal, GSM : Orfanidou et al., 2006 ; au niveau du gyrus angulaire, GA : Orfanidou et al., 2006 ; Raettig & Kotz, 2007).

Ces régions pourraient être impliquées dans le décodage lexico-sémantique lors du processus de reconnaissance de mots isolés. Etroitement relié à ce propos, un travail ayant utilisé l'effet Ganong (cf. section 2.2.5), suggère que l'influence du lexique sur le processus de catégorisation phonétique couramment observé avec ce paradigme pourrait être justement localisé au carrefour temporo-pariétal, au niveau du GSM (Gow, Segawa, Ahlfors, & Lin, 2008) ou dans la partie supérieure du lobe temporal, au niveau du gyrus temporal supérieur, GTS (Myers & Blumstein, 2008). Ajoutons que d'autres travaux supposent que les régions situées autour de la partie postérieure du GTM et du STS pourraient jouer le rôle d'interface entre l'arrivée d'un stimulus sensoriel et son appariement avec une représentation abstraite en mémoire (e.g., Gagnepain et al., 2008; Hickok & Poeppel, 2007).

En conclusion, suite à l'examen de ces différentes études, il semblerait que le lobe temporal ainsi que la partie inférieure du lobe pariétal jouent un rôle important dans le processus d'accès au lexique.

2.5. CONCLUSIONS

Dans ce chapitre, nous avons passé en revue différentes études et modèles décrivant le processus de reconnaissance de mots. A travers l'examen de divers arguments issus de l'étude des caractéristiques du signal acoustique de parole, nous avons expliqué pourquoi la majorité des modèles d'accès au lexique en modalité auditive postule que tout individu adulte dispose de représentations abstraites de mots en mémoire. Nous avons également évoqué des résultats comportementaux et des travaux effectués en neuro-imagerie, permettant de justifier la distinction entre niveau lexical et pré-lexical dans le processus de traitement de la parole. Nous nous sommes ensuite focalisés sur trois questions importantes étudiant le processus de reconnaissance de mots en modalité auditive : le format des unités permettant de contacter le lexique, le décours temporel du traitement du signal acoustique, mais également l'influence du statut lexical sur le décodage du signal de parole, ces dernières étant reliées à l'étude de notre problématique. Enfin, nous avons présenté plusieurs modèles psycholinguistiques décrivant le processus d'accès au lexique en modalité auditive seule, s'opposant notamment sur les questions évoquées plus haut. C'est dans le cadre de ces modèles que nous interprèterons les données issues de différentes études expérimentales présentées dans ce manuscrit, visant à évaluer le rôle de l'information visuelle dans le processus d'accès au lexique.

2.6. PROBLEMATIQUE GENERALE

Dans les parties précédentes nous avons présenté divers travaux et modèles ayant étudié *séparément* l'influence de l'information visuelle (Chapitre 1) et lexicale (Chapitre 2) dans le processus de perception de la parole. L'objectif de ce travail de thèse consiste à étudier de manière *combinée* le rôle de ces deux types d'informations dans le processus de reconnaissance de mots. Plus précisément, le but de nos recherches est de déterminer si le signal visuel de parole contribue au processus de reconnaissance de mots per se, autrement dit dans le décodage de l'information lexicale.

La question du rôle de l'information visuelle dans le processus de reconnaissance de mots peut, de prime abord, paraître triviale. En effet, certains mots peuvent être « lus sur les lèvres ». Considérons la situation de communication classique où un locuteur A est dans l'impossibilité d'émettre le moindre son, mais doit néanmoins transmettre un message oral à son interlocuteur B. Dans ce cas, si A articule silencieusement et suffisamment distinctement un mot et si ce dernier est aisément prédictible par rapport au contexte de la situation de communication, il y a de fortes chances pour que B comprenne ce qui vient d'être dit, alors

même qu'aucun son n'est sorti de la bouche de A. En conséquence, il est probable que l'interlocuteur B parvienne à percevoir le message que veut lui transmettre A, en l'absence de toute information auditive. Ainsi, l'examen superficiel de cette situation pourrait nous conduire à la conclusion prématurée que le fait de voir les gestes articulatoires de son interlocuteur nous permet directement de reconnaître des mots, autrement dit d'en extraire du sens. Cependant, dans cette situation précise, en plus de disposer d'indices visuels exagérés liés à l'hyper-articulation de A, l'interlocuteur B peut utiliser des indices contextuels spécifiques à la conversation. Ainsi, il est tout à fait possible que B se soit servi d'informations contextuelles pour deviner ce qui a été dit. La question clé de ce travail de recherche consiste à déterminer si le fait de voir le visage de son interlocuteur permet de contacter *automatiquement* les représentations de mots contenues dans notre lexique mental.

Par conséquent, la problématique générale ayant motivé nos recherches est de déterminer si le signal visuel de parole permet d'activer les unités lexicales. Pour cela, rappelons que dans le Chapitre 2, nous avons vu que la majorité des modèles décrivant l'accès au lexique en modalité auditive supposent globalement deux grandes étapes du décodage de la parole : un niveau pré-lexical et un niveau lexical. Dans le Chapitre 1, nous avons évoqué un grand nombre d'études indiquant que le fait de voir le visage de son interlocuteur permet d'améliorer la perception des sons de parole (e.g., lorsque celui-ci est détérioré par du bruit : Sumby & Pollack, 1954). Ce premier constat montre que l'information extraite de la gestualité oro-faciale aide à décoder le signal acoustique de parole, au moins à un niveau pré-lexical. Cependant, ces données ne nous permettent pas de déterminer si l'information visuelle permet d'intervenir à un niveau lexical de décodage de la parole. En effet, rappelons que la mise en action de certains articulateurs n'est pas visible (e.g., vibration des cordes vocales) ce qui rend la perception de certains traits phonétiques impossible en modalité visuelle seule (e.g., le voisement). Cela a pour conséquence qu'un même geste articulatoire peut, en l'absence de toute information contextuelle, correspondre à plusieurs mots, donc à plusieurs représentations différentes (e.g., « bain » vs. « pain »). Déterminer si le traitement de cette information visuelle permet d'activer les représentations lexicales, alors que seul un très faible pourcentage de mots peut être consciemment identifié sur la seule base du geste articulatoire de parole (10 à 20 %, Schwartz, 2011), constitue donc une véritable question de recherche, dont la réponse est loin d'être évidente.

L'objet des prochains chapitres expérimentaux consiste à explorer ce phénomène chez l'adulte, en modalité audiovisuelle (Chapitre 3) et visuelle seule (Chapitre 4) ainsi que chez l'enfant, en situation audiovisuelle (Chapitre 5).

CHAPITRE 3. APPORT DE L'INFORMATION VISUELLE DANS L'ACCES AU LEXIQUE EN SITUATION BRUITEE

3.1. INTRODUCTION

Dans ce chapitre, nous allons spécifiquement nous intéresser au rôle de l'information visuelle dans le processus d'accès au lexique chez l'adulte, en modalité audiovisuelle. De ce fait, nous allons donc passer en revue différentes études ayant étudié cette question exclusivement en modalité audiovisuelle, avant de présenter nos propres travaux à ce sujet.

3.1.1. Travaux antérieurs

A notre connaissance, Trout et Poser (Trout & Poser, 1990) ont été parmis les premiers à avoir effectués des travaux étudiant l'impact de l'information visuelle et lexicale, en présence d'une information auditive. Pour cela, ces auteurs ont utilisé un paradigme de restauration phonémique. Initialement découvert en modalité auditive (e.g., Samuel, 1981; Warren, 1970) ce paradigme met en évidence que lorsqu'une portion acoustique d'un mot (e.g., un phonème) est remplacée par du bruit, ce signal de parole est néanmoins perçu comme intact, on dit que le segment manquant dans le signal est restauré perceptivement. Cette illusion perceptive étant observée plus fréquemment dans des mots que dans des pseudo-mots (Samuel, 1981, 1996) (cf. Chapitre 2, section 2.2.5) cette dernière peut être (en partie) expliquée par l'influence du contexte lexical sur la perception du signal de parole. Ainsi, dans l'étude de Trout et Poser (1990), un segment du signal acoustique d'une phrase (e.g., « Government forces put down the student re**b**ellion[29] », « les forces gouvernementales ont stoppé la rébellion étudiante ») était soit remplacé par du bruit (condition « remplacé »), soit du bruit était ajouté au niveau de ce segment, laissant le signal acoustique intact (condition « ajouté »). Chacune de ces phrases était présentée en modalité auditive ou audiovisuelle. La tâche des participants était de désigner si dans la phrase perçue, le bruit était présenté en condition « ajouté » ou « remplacé ». Selon l'hypothèse que (1) l'effet de restauration phonémique est en partie dû à une influence du niveau lexical (2) l'information visuelle permet d'activer les représentations de mots dans le lexique, les auteurs auraient dû observer un effet de restauration phonémique plus fréquent en modalité audiovisuelle qu'auditive seule. Cependant, leurs résultats montrent au contraire que la présence de l'information visuelle avait pour impact de *réduire* l'illusion de restauration phonémique. Ces données, quelques peu surprenantes, ont poussé les auteurs à invoquer des biais méthodologiques permettant de justifier que l'*ajout* d'une information visuelle congruente relative au phonème manipulé ait un impact négatif sur la perception du signal de parole comme intact.

[29] La lettre en gras représente le phonème sur lequel porte l'illusion

Presque vingt ans plus tard, dans une étude similaire, Shanin et Miller (Shahin & Miller, 2009) ont trouvé des résultats différents. En utilisant également un paradigme de restauration phonémique, ces derniers ont observé que cet effet était plus important en modalité audiovisuelle qu'auditive seule pour des mots trisyllabiques isolés. Egalement, ces auteurs ont mis en évidence que ce bénéfice lié à la présence de l'information visuelle était uniquement observé lors de la présentation d'un visage en *mouvement* articulant un signal *congruent* avec l'information auditive, mais pas lors de la simple représentation d'un visage *statique* ou articulant une information *non congruente* avec le signal acoustique. Cette différence indique que le bénéfice lié à la présence du signal visuel de parole ne pouvait pas être expliquée en termes de facteurs purement attentionnels pouvant distraire/focaliser l'attention des participants et ainsi moduler l'effet de restauration phonémique. Ainsi, contrairement aux résultats de Trout et Poser (1990), ces données pourraient suggérer que l'information visuelle permet d'activer les représentations lexicales. Cependant, il a été montré que l'effet de restauration phonémique est dû à l'influence conjointe du niveau lexical mais est aussi sensible à des paramètres liés au signal d'entrée (e.g., Shahin & Miller, 2009; Trout & Poser, 1990). Les auteurs n'ayant pas utilisé de pseudo-mots dans leur étude, aucune conclusion claire ne peut être tirée de leurs résultats quant au locus du bénéfice observé par la présence de l'information visuelle.

Certains travaux ont étudié l'impact de paramètres lexicaux sur la prise en compte de l'information visuelle en présence du signal acoustique de parole (e.g., Kaiser, Kirk, Lachs, & Pisoni, 2003; Tye-Murray, Sommers, & Spehar, 2007). Kaiser et al. (2003), se sont par exemple intéressés à l'influence de la fréquence lexicale et de la densité du voisinage phonologique (cf. Chapitre 2 section 2.3.5) sur la perception de la parole en modalité auditive et audiovisuelle. Pour cela, les auteurs ont créé deux catégories de stimuli : les mots « faciles » à reconnaître, c'est-à-dire fréquents dans le langage oral et disposant de peu de voisins opposés aux mots « difficiles » à reconnaître, c'est-à-dire peu fréquent et ayant une grand nombre de voisins. Les participants avaient pour tâche d'identifier un mot dont le signal acoustique était détérioré par du bruit (RSB = -5 dB). Les stimuli étaient présentés soit en modalité auditive, soit en modalité audiovisuelle[30]. L'idée sous-tendue par cette manipulation est que si l'information visuelle est décodée à un niveau lexical, des facteurs inhérents à ce niveau de traitement devrait influencer son décodage. En effet, la fréquence lexicale, qui correspond à la fréquence d'occurrence d'un mot dans le langage oral mais également la nombre de voisins phonologiques, qui désigne, selon le modèle NAM, le

[30] Notons que les performances des participants ont également été évaluées en modalité visuelle seule. Les effets obtenus dans cette condition seront reportés dans le chapitre suivant.

nombre de compétiteurs activés lors des premières phases de l'accès au lexique (voir Luce & Pisoni, 1998, pour une revue), sont des caractéristiques spécifiques du niveau lexical. Le sens des prédictions formulées par les auteurs était que le bénéfice lié à la présence de l'information visuelle, lorsque l'information auditive est détériorée, devrait être plus important pour les mots « faciles » que « difficiles » à reconnaître. Cependant, bien que leurs résultats mettent en évidence de meilleures performances en modalité audiovisuelle qu'auditive seule, seule une influence tendancielle (i.e., non significative sur le plan statistique) de ces paramètres lexicaux sur le bénéfice lié à la présence de l'information visuelle a été mise en évidence. Ainsi, ces résultats suggèrent que l'information visuelle serait décodée à un niveau lexical mais n'apportent pas de preuve tangible vis-à-vis de cette hypothèse.

L'étude de Tye-Murray et al. (2007) a pour sa part évalué uniquement l'impact de la densité du voisinage phonologique sur le processus de reconnaissance de mots en modalité audiovisuelle. L'originalité de ce travail réside dans le fait que la densité du voisinage phonologique a été évaluée séparément pour la modalité auditive et visuelle. Seule l'intersection de ces deux évaluations séparées a été retenue comme voisinage des stimuli présentés en modalité audiovisuelle. Ainsi, le voisinage « acoustique » de chaque stimulus a été déterminé selon la règle établie par Luce et Pisoni (1998) : par exemple, le mot /fɔːk/, « fork », fourchette, dispose de 13 voisins, par addition, substitution ou délétion d'un phonème : « forked », « force » « for », etc. La densité du voisinage « visuelle » a été établie en ne sélectionnant que les mots dont l'ensemble des phonèmes consonantiques correspondent aux mêmes visèmes (e.g., pour « fork », sont considérés comme voisins « visuels » les mots « force », « farce », etc.). Seuls les items étant considérés à la fois comme voisins « acoustiques » et « visuels » ont été retenus pour définir le voisinage audiovisuel de chacun des stimuli. La tâche proposée aux participants était d'identifer le mot présenté en modalité audiovisuelle le signal acoustique des stimuli était détérioré par du bruit. Les résultats ont mis en évidence que lorsque les mots étaient présenté en modalité audiovisuelle, que la densité du voisinage « acoustique » et « visuel » avaient tous les deux une influence sur les performances des participants. Cela suggère donc que lors de la perception d'un mot en modalité audiovisuelle, l'ensemble des candidats lexicaux activés est compatible à la fois avec le signal acoustique et mais également visuel. En d'autres termes, les résultats de cette étude suggèrent que l'information visuelle permet de contacter les unités lexicales, en présence d'une information auditive détériorée par du bruit.

Remarquons cependant que l'ensemble des travaux que nous venons de passer en revue obtiennent des conclusions assez différentes : alors que deux études pourraient

éventuellement indiquer que le signal visuel de parole serait décodé à un niveau lexical (i.e., Shahin & Miller, 2009; Tye-Murray, et al., 2007) les deux autres n'ont pas réussi à obtenir des résultats allant dans le sens de cette interprétation (i.e., Kaiser, et al., 2003; Trout & Poser, 1990).

A notre connaissance, quatre études se sont intéressées à l'apport de l'information visuelle dans l'accès au lexique en modalité audiovisuelle, en opposant les performances obtenues avec des mots avec celles obtenues avec du matériel non lexical comme des pseudo-mots ou des non-mots (Barutchu, Crewther, Kiely, Murphy, & Crewther, 2008; Brancazio, 2004; Sams, 1998; Windmann, 2004). Notons également que pour la totalité de ces travaux, cette question a été étudiée en présence d'une information auditive non détériorée et non congruente avec le geste articulatoire. En effet, le paradigme utilisé était celui de l'illusion McGurk (McGurk & McDonald, 1976, cf. Chapitre 1, section 1.3.2.2). Rappelons que cet effet, mainte fois répliqué dans la littérature (voir Colin & Radeau, 2003, pour une revue) a mis en évidence qu'un signal acoustique /ba/ doublé de l'articulation d'un /ga/ est souvent perçu comme /da/ ou /ða/. Cette illusion est la preuve que les informations auditives et articulatoires sont intégrées ou fusionnées par notre système perceptif.

Ainsi, dans une étude menée en finnois, Sams et al. (1998) ont comparé le pourcentage de percept correspondant à une intégration audiovisuelle dans deux situations. Dans une première condition, le signal acoustique d'un mot (e.g., /**p**anu[31]/ « **p**annu » qui signifie. « poêle »), doublé de l'articulation d'un autre mot (e.g., /**k**anu/, « **k**annu » qui signifie « cruche ») était présenté aux participants, le percept attendu étant un pseudo-mot (e.g., /**t**anu/). Dans une deuxième condition, deux pseudo-mots différents (e.g., /**p**iili/ en modalité auditive et /**k**iili/ dans le signal visuel) étaient présentés, donnant lieu à la perception du mot « tiili », /**t**iili/ (signifiant « brique »). Faisant l'hypothèse que l'information visuelle contribue au processus de reconnaissance de mot (et donc à l'activation des représentations lexicales) les auteurs s'attendaient à obtenir un pourcentage d'effet McGurk plus important (et donc une prise en compte des informations visuelles plus importante) lorsque l'intégration audiovisuelle donnait lieu à la perception d'un mot plutôt que d'un pseudo-mot. Cependant, leurs résultats se révélèrent aller dans le sens contraire à leurs attentes. En effet, seulement 36% des réponses correspondant à l'intégration des informations auditives et visuelles ont donné lieu à la perception d'un mot (e.g., « **t**iili ») alors qu'un effet McGurk a été observé dans 50% des cas pour la perception d'un pseudo-mot (e.g.,/**t**anu/). Sur la base

[31] Les phonèmes non congruents entre l'information visuelle et auditive et sur lesquels l'intégration de ces deux signaux est censée porter sont représentés en gras dans l'ensemble des exemples.

de ces résultats, les auteurs ont conclu que l'information visuelle n'influençait pas l'activation des unités lexicales.

Cependant, certaines critiques ont été effectuées envers ce travail (Brancazio, 1999, 2004). En effet, il semblerait que certaines variables pouvant moduler l'effet McGurk n'aient pas été contrôlées dans l'étude de Sams et al. (1998). Tout d'abord, la position du phonème consonantique non congruent n'était pas la même dans toutes les conditions de présentation des items. Ensuite, la nature de la voyelle succédant à ce phonème pouvait varier entre les différentes conditions, alors qu'il s'agit d'un facteur connu pour influencer la taille de l'effet McGurk (Green, Kuhl, Meltzoff, & Stevens, 1991). Dans une série d'études menées en anglais, Brancazio (Brancazio, 1999, 2004) a également étudié cette question tout en contrôlant les variables de l'étude de Sams et al. (1998). Dans cette étude, une version spécifique de l'effet McGurk a été utilisée. Cette dernière consiste à présenter la syllabe /ba/ dans le signal acoustique doublée de l'articulation d'un /da/ (et non d'un /ga/ comme dans l'étude princeps de McGurk & MacDonald, 1976). Dans cette version de l'illusion, le percept attendu est un /da/, la modalité visuelle dominant alors complètement le percept (e.g., Repp, Manuel, Liberman, & Studdert-Kennedy, 1983, cité dans Colin & Radeau, 2003). En combinant le paradigme de Ganong (Ganong, 1980, cf. Chapitre 2 section 2 2.3) et l'effet McGurk, Brancazio a proposé une tâche de catégorisation de phonèmes dans des stimuli présentés en modalité audiovisuelle, selon deux conditions différentes. Dans une première condition, un mot était contenu dans le signal acoustique (e.g., /bɛg/, « beg », « mendier ») et doublé avec l'articulation d'un pseudo-mot (e.g., /dɛg/). Dans une seconde condition, le mot était présenté dans le signal visuel (e.g., /dɛsk/, « desk », « bureau ») et accompagné d'un pseudo-mot dans le signal acoustique (e.g., /bɛsk/). La tâche consistait à catégoriser le premier phonème (soit /b/ soit /d/). Les résultats montrent tout d'abord que la perception du phonème était biaisée en faveur du contexte lexical. Pour l'ensemble des conditions, le pourcentage d'identification d'un son consonantique était plus élevé lorsque celui-ci était contenu dans un mot que dans un pseudo-mot (cf. Ganong, 1980). Brancazio a observé que cet effet lexical était plus important lorsque le mot était contenu dans le signal visuel que dans le signal auditif. Ce résultat suggère donc que le contexte lexical influence non seulement la prise en compte des informations contenues dans le signal auditif (Ganong, 1980) mais également la prise en compte des informations visuelles lors du processus de reconnaissance de mots. Avec un paradigme similaire Barutchu et collègues (Barutchu, et al., 2008) ont répliqué les résultats trouvés par Brancazio (Brancazio, 1999, 2004). De plus ils ont montré que la place du phonème sur lequel porte l'illusion modifiait le pourcentage

d'effet McGurk obtenu sur les mots, mais pas celui obtenu sur les pseudo-mots. Plus précisément, ces auteurs ont mis en évidence que le signal acoustique d'un mot présenté avec le signal visuel d'un autre mot, donnait plus fréquemment lieu à la perception illusoire d'un troisième mot lorsque le phonème sur lequel portait l'effet était situé en début (e.g., /**b**ɪl/, « **b**ill », facture, en auditif + /**g**ɪl/, « gill », branchie, en visuel → /**d**ɪl/, « dill », aneth) plutôt qu'en fin de mot (e.g., /la**b**/, « la**b** », laboratoire, en auditif + /la**g**/, « la**g** », décalage, en visuel, → /la**d**/, « la**d** », palefrenier). Aucune différence de la sorte n'ayant été obtenue sur les pseudo-mots, les auteurs en ont conclu que le processus d'intégration des informations visuelle et auditive est un mécanisme s'effectuerait lors d'une étape précédant au le processus d'identification de mots.

Enfin, une étude effectuée en allemand par Windmann (Windmann, 2004) a apporté des résultats similaires à ceux de Brancazio (2004) et Barutchu et al. (2008). De la même manière que Sams et al. (1998), Brancazio (2004) et Barutchu et al. (2008), l'effet McGurk était utilisé en tant que paradigme expérimental dans cette étude. La particularité de ce travail par rapport aux études précédentes est que les stimuli proposés étaient présentés isolément mais étaient précédés par la présentation écrite d'une phrase. Ces stimuli pouvaient soit être congruents sémantiquement avec cet énoncé, soit non congruent. Par exemple, dans la condition congruente, le stimulus correspondant à l'illusion perceptive de /zuker/, « zucker », sucre, résultant de la présentation auditive et visuelle de deux pseudo-mots (e.g., /zuper/ + /zuter/) était présenté après la phrase signifiant « Je préfère prendre mon café avec du lait et du… ». Dans la condition non congruente, la présentation auditive et visuelle de deux pseudo-mots (i.e., /glɔp/ + /glɔt/, respectivement) donnant lieu à la perception du stimulus /glɔk/, « Glocke », montre, était proposé après cette même phrase. La tâche des participants était de répéter le stimulus perçu. Leurs résultats mettent en évidence un effet McGurk plus important (et donc une prise en compte des informations visuelle plus importante) lorsque le résultat de cette illusion présenté était congruent avec le contexte lexico-sémantique de la phrase présentée au préalable. Ainsi, ces résultats suggèrent que le contexte lexico-sémantique influence la perception du signal visuel de parole. Cette influence pouvant être due à l'activation automatique des représentations lexicales, mais également au développement d'attentes conscientes vis-à-vis du stimulus, Windmann (2004) a dans une seconde expérience, manipulé les attentes perceptives vis-à-vis de la nature d'un stimulus non significatif. Dans cette logique, si l'effet observé est uniquement due à des stratégies conscientes, le même biais contextuel devrait être obtenu avec des non-mots. Pour cela, le signal acoustique d'un non-mot de type VCV (e.g., /aba/)

était doublé de l'articulation d'un autre non-mot (e.g., /aga/) donnant lieu à la perception illusoire d'un troisième non-mot (e.g., /ada/). Un non-mot de type VCV permettant de manipuler les attentes perceptives des participants était présentée à l'écrit avant chaque stimulus. Il pouvait être soit congruent avec le signal acoustique (e.g., /aba/), soit le signal visuel (e.g., /aga/), soit avec le percept illusoire attendu (e.g., /ada/). Les participants avaient ensuite pour tâche de comparer ensuite le stimulus perçu avec le non-mot présenté à l'écrit. Bien que les résultats montrent un pourcentage d'effet McGurk plus important lorsque le non-mot présenté à l'écrit était congruent avec cette illusion (e.g., présentation écrite de /ada/), cette influence s'est révélée beaucoup plus faible que dans l'expérience précédente, lorsque des mots étaient présentés. En considérant les résultats issus de ces deux expériences, ces résultats suggèrent que le décodage de l'information visuelle est influencé par le contexte lexico-sémantique. Par conséquent, cette étude suggère donc que le signal visuel de parole contribue à un décodage de l'information à un niveau lexical.

En conclusion, nous pouvons premièrement constater que l'examen de l'ensemble de ces études n'est pas unanime vis-à-vis de la question étudiée. En effet, trois (i.e., Kaiser, et al., 2003; Sams, 1998; Trout & Poser, 1990) obtiennent néanmoins des résultats mitigés, alors que cinq études suggèrent que l'information participe au processus d'activation des unités lexicales (i.e., Barutchu, et al., 2008; Brancazio, 1999, 2004; Shahin & Miller, 2009; Tye-Murray, et al., 2007; Windmann, 2004). Ensuite, dans ce dernier groupe, la totalité des travaux présentés ont utilisé l'effet McGurk. Bien que celui-ci constitue un outil intéressant pour l'étude de la perception de la parole en modalité audiovisuelle cette illusion ne résulte pas moins d'une modification artificielle du langage oral. Or cette situation d'incongruence entre le signal visuel et acoustique de parole est rarement observée dans la vie quotidienne (excepté lors de la perception de films doublés). En effet, à la différence des illusions obtenues pour la perception des objets en modalité visuelle, celle-ci n'est expérimentée qu'en situation de laboratoire. La validité écologique des conclusions effectuées sur la base des études ayant utilisé cette illusion perceptive comme paradigme est donc grandement amoindrie. Enfin, la totalité de ces travaux fournit des arguments récoltés sur les pourcentages de réponses correctes qui sont mesurés en *temps différé*, c'est-à-dire après la présentation du stimulus, lorsqu'aucune contrainte temporelle de réponse n'est donnée aux participants. Or aucune de ces études ne présente de données obtenues sur les temps de réponse, permettant d'obtenir une mesure sensible en *temps réel* des processus impliqués. L'objectif des Etudes 1 et 2 de ce chapitre consiste à étudier le rôle de l'information visuelle dans le processus d'accès au lexique, en palliant aux limites des études citées plus haut.

3.1.2. Objectifs et méthodes des Etudes 1 et 2

Le but des Etudes 1 et 2 présentées dans ce chapitre consiste à tester si la perception des mouvements oro-faciaux d'un interlocuteur contribue au processus d'activation des représentations lexicales. L'originalité de ces deux premières études est d'aborder cette question en *présence* d'une information auditive détériorée par du bruit, mais *congruente* avec le signal visuel. A cette fin, des tâches de détection de phonèmes consonantiques (Etude 1) et vocaliques (Etude 2) ont été utilisées. Le paradigme de détection de phonèmes a tout d'abord été sélectionné car il permet d'étudier (notamment) l'influence de l'information lexicale sur la perception des phonèmes ; elle a initialement été utilisée en modalité auditive (voir Connine & Titone, 1996, pour une revue de question). Classiquement, la détection d'un phonème est plus rapide lorsque celui-ci est inséré dans un mot plutôt que dans un pseudo-mot (i.e., effet de supériorité du mot, e.g., Frauenfelder et al., 1990, voir Chapitre 2, section 2.2.5 pour plus de détails à ce sujet). Ce résultat indique que l'information lexicale influence le processus de traitement des phonèmes en modalité auditive. Une seconde raison pour laquelle nous avons choisi d'utiliser ce paradigme vient du fait que la tâche de détection de phonèmes permet de fournir, en plus du nombre de réponses correctes, une mesure des temps de réponse de chaque participant, permettant d'évaluer en temps réel l'apport de l'information visuelle au processus de reconnaissance de mots. En conséquence, cette tâche devrait nous permettre de fournir des données supplémentaires par rapport aux paradigmes évoqués précédemment.

Afin d'adapter ce paradigme à l'étude spécifique de notre problématique, cette tâche de détection de phonèmes a été effectuée dans des mots et des pseudo-mots, présentés en modalité Auditive (A) ou Audiovisuelle (AV). L'objectif d'opposer les performances obtenues avec les mots et les pseudo-mots (et non avec des non-mots) était d'examiner l'implication du niveau lexical sur le traitement des phonèmes en prenant pour condition contrôle des items non lexicaux tout en respectant les contraintes phonotactiques de la langue française. La présentation des stimuli était effectuée avec différents niveaux de détérioration du signal acoustique (i.e., Rapport Signal Sur Bruit), afin (1) d'éviter un plafonnement des performances pour l'ensemble des conditions et (2) de maximiser l'apport de l'information visuelle en présence de l'information auditive. En effet, cette détérioration du signal acoustique devrait également faciliter la mise en évidence d'un bénéfice lié à la présence de l'information visuelle en présence du signal acoustique (AV) par rapport à une situation où seule l'information auditive (A) est disponible (e.g., Benoît, et al., 1994). Si, conformément à nos hypothèses, l'information visuelle relative aux gestes articulatoires d'un interlocuteur contribue au processus d'activation des représentations lexicales, nous devrions obtenir un effet de supériorité du mot plus important en condition AV qu'en condition A.

3.2. ETUDE 1 : INFLUENCE DE L'INFORMATION VISUELLE ET LEXICALE DANS LE PROCESSUS DE DETECTION DE PHONEMES CONSONANTIQUES

Fort, M., Spinelli, E., Savariaux, C. & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*. 52 (6), 525-532.

3.2.1. Méthode

3.2.1.1. Participants

Quatre-vingt-un participants (dont 59 femmes et 22 hommes) âgés de 18 à 51 ans ($M = 23$ ans) ont été recrutés pour cette étude. Tous étaient de langue maternelle française. Aucun d'entre eux n'avait de déficit sur le plan auditif et tous avaient une vision normale ou corrigée. La majorité d'entre eux étaient étudiants en Psychologie à l'Université Pierre Mendès France de Grenoble et recevaient un bon d'expérimentation en échange de leur participation.

3.2.1.2. Stimuli

3.2.1.2.1. Items expérimentaux

Un corpus de 34 paires de mots/pseudo-mots bisyllabiques de type CVCV a été sélectionné en utilisant la base de données LEXIQUE 2 (New, Pallier, Ferrand, & Matos, 2001, cf. Annexe A). Chacun de ces items comportait un des sept phonèmes-cibles consonantiques sélectionnés pour cette étude : trois de ces phonèmes-cibles présentant un lieu d'articulation labial ou labiodental (/p/, /f/, /v/) et quatre présentant un lieu d'articulation dental ou alvéolaire (/d/, /t/, /s/, /z/). Notons ici que certains phonèmes-cibles avaient tous un lieu d'articulation hautement visible, ces derniers étant articulés à l'avant du conduit vocal (au niveau des lèvres pour les labiales). Le phonème-cible était toujours situé au début de la seconde syllabe et chaque membre de la paire mot/pseudo-mot était identique excepté pour la voyelle finale (e.g., le phonème-cible /**p**/ dans « cha**p**eau » ou /ʃapy/). Le phonème-cible était également situé à la fin du mot plutôt qu'au début, la probabilité d'observer un effet lexical robuste étant plus importante dans ce cas (Frauenfelder, et al., 1990). Pour chaque stimulus, le phonème vocalique final pouvait engendrer soit un mouvement d'arrondissement (/o/, /u/, /y/, /œ/) soit un mouvement d'étirement des lèvres (/i/, /e/). En conséquence, la moitié des paires mot/pseudo-mot était « contrastée » pour le geste

articulatoire de ce dernier phonème. Dans cette condition, lorsqu'un membre de celle-ci comportait une voyelle finale arrondie, l'autre membre se terminait par une voyelle étirée (e.g., /tʁupo/, « troupeau » vs. /tʁupi/). Pour la condition « non contrastée », les deux membres de chaque paire se terminaient toujours par le même type de phonème (e.g., /ʃapo/ vs. /ʃapy/). La fréquence lexicale moyenne des mots porteurs du phonème-cible était de 45.88 occurrences par million (opm) dans le langage oral. La base de données « Freqfilm2 », de LEXIQUE 2 (New, et al., 2001) a été utilisée pour effectuer cette estimation. La fréquence lexicale des mots dans le langage oral étant connue pour influencer la reconnaissance des mots en modalité visuelle (e.g., Mattys, et al., 2002), nous avons contrôlé ce facteur. La moitié de ces items était considérée comme relativement fréquente dans le langage oral (Fréquence F > 10 opm) alors que l'autre moitié ne l'était pas (F < 10 opm). La durée d'élocution moyenne à partir du début du phonème-cible pour les mots ne différait pas de celle des pseudo-mots (M_{mots} = 554 ms ; $M_{pseudo\text{-}mots}$ = 551 ms, $t(67) < 1$).

3.2.1.2.2. Items de remplissage

Afin que les participants ne répondent pas que le phonème-cible est présent à chaque essai, 40 paires de mot/pseudo-mot correspondant aux items de remplissage (i.e., ne contenant pas le phonème-cible à détecter) ont été construits en utilisant les mêmes caractéristiques que les items expérimentaux. Cependant, à la différence de ces derniers, les items de remplissage étaient toujours associés à un phonème différent ne correspondant à aucun de ceux contenus dans l'item en question (e.g., le phonème-cible /**p**/, pour /toʁty/, « tortue » vs. /toʁti/). Ces phonèmes de remplissage pouvaient différer d'une ou deux caractéristiques (ou traits) articulatoires (e.g., place d'articulation et/ou mode d'articulation) par rapport aux phonèmes contenus dans l'item cible. La fréquence moyenne de ces items était de 42.99 opm. La moitié de ces items était considérée comme fréquente (Fréquence F > 10 opm) alors que l'autre moitié ne l'était pas (F < 10 opm).

3.2.1.2.3. Enregistrement des stimuli

L'ensemble des items expérimentaux et de remplissage a été enregistré dans une chambre sourde à l'aide d'un microphone AKG C1000S et d'une caméra vidéo tri-CCD SONY DXC-990P. Ils ont été prononcés par un locuteur entraîné, de langue maternelle française. Ce dernier était placé devant un fond vert et seuls sa tête, son cou et le haut de ses épaules étaient visibles (cf. Figure 24). Il avait pour consigne d'initier la production de chaque item en partant avec la bouche fermée et de ne pas cligner des yeux durant la prononciation de chacun d'eux. Les stimuli ont été numérisés à l'aide du logiciel Dps Reality v 3.1.9 pour obtenir des fichiers vidéo compressés au format mpeg2.

Les phonèmes-cibles ont également été enregistrés dans une chambre sourde à l'aide d'un microphone AKG C1000S et d'un enregistreur numérique Marantz PMD 670 afin d'obtenir des fichiers audio au format WAV. Ils ont été prononcés par une locutrice de 22 ans de langue maternelle française dans le contexte vocalique d'un schwa (e.g., le phonème-cible /pə/ à détecter dans /ʃapo/ ou /ʃapy/). Deux locuteurs différents ont donc été choisis pour prononcer le matériel utilisé dans cette expérience afin de s'assurer que l'identité du phonème et non une caractéristique spécifique du locuteur était utilisée pour effectuer la tâche.

Le Rapport Signal sur Bruit (RSB) de chaque item a été calculé et modifié en générant du bruit blanc à l'aide du logiciel Matlab. Le Rapport Signal sur Bruit, exprimé en Décibel (dB), désigne le rapport entre la puissance du signal auditif et la puissance des bruits parasites comme par exemple le bruit de fond. Pour diminuer ce rapport en fonction des conditions de présentation souhaitées, du bruit blanc a été ajouté à chaque item, en utilisant la formule suivante : $RSB = 20 \log_{10}(\text{Signal A}/\text{Bruit A})$ où Signal A et Bruit A sont respectivement les amplitudes du signal original et du bruit. Les stimuli n'ayant pas la même énergie au cours du temps, celle-ci dépendant des propriétés acoustiques des différentes voyelles et consonnes de chaque item, chaque RSB a été calculé à partir de la puissance moyenne de chaque stimulus correspondant. Dans cette étude, deux niveaux de bruit ou RSB ont été utilisés : -9 dB versus -18 dB. Cette permet d'étudier l'apport de l'information visuelle au processus d'accès au lexique selon deux niveaux de détérioration de l'information auditive. En effet, ces niveaux de bruits ont respectivement été choisis afin d'obtenir (1) une condition où le signal acoustique est détérioré (-9 dB) mais reste relativement intelligible en l'absence de l'information visuelle (e.g., dans approximativement 50 % des cas, Benoît, et al., 1994; Binnie, et al., 1974) ; (2) une condition où le RSB (-18 dB), permet une identification correctes de phonèmes vocaliques et consonantiques en modalité auditive seule faible mais possible (e.g., entre 0 et 10 %, Benoît, et al., 1994; Binnie, et al., 1974). Nous avons

sélectionné ces niveaux de bruit à dessein comme comportant une valeur de RSB « intermédiaire ». Cela avait pour but d'éviter des situations extrêmes correspondant à une très faible (e.g., à 0 dB) ou à une très forte détérioration du signal acoustique (e.g., -24 dB), afin de maximiser l'apport de la modalité audiovisuelle par rapport à une situation auditive seule (Ross, et al., 2007).

Chaque paire de mot/pseudo-mot était répartie selon 4 listes correspondant aux 4 conditions de présentation des stimuli : auditive seule à -9 dB ; auditive seule à -18 dB ; audiovisuelle à -9 dB ; audiovisuelle à -18 dB. Chaque liste contenait 8 ou 9 paires d'items expérimentaux et 10 paires d'items de remplissage. Chaque item (expérimental ou de remplissage) était présenté une seule fois à chaque participant. Chaque phonème-cible était présenté un nombre égal de fois tout au long de l'expérience. La modalité de présentation de chaque liste était contrebalancée entre les participants.

3.2.1.3. Procédure

Les participants étaient évalués individuellement. Ils étaient assis dans une chambre sourde à 50 cm d'un écran LCD de 17 pouces (Neovo 17 X-17A). La composante vidéo des stimuli était présentée à une fréquence de 25 images/seconde. Le signal acoustique était présenté à une fréquence d'échantillonnage de 44100 Hz par deux enceintes de type SONY SRS-88 situées de chaque côté de l'écran. La consigne était donnée à l'oral par l'expérimentatrice. Il leur était expliqué qu'ils allaient entendre le phonème-cible, suivi d'un mot ou d'un pseudo-mot (i.e., le stimulus-cible) contenant - ou ne contenant pas - le phonème à détecter.

Avant chaque essai, le participant voyait un point de fixation. Ensuite, le phonème-cible était présenté en modalité auditive seule alors qu'un écran noir s'affichait simultanément. Un second point de fixation[32] précédait toujours le stimulus-cible. Ce dernier pouvait être présenté en modalité auditive seule (A) avec un fond noir à l'écran ou en modalité Audiovisuelle (AV, i.e., avec la vidéo du locuteur). Les participants devaient appuyer le plus rapidement possible (en utilisant la même main tout au long de l'expérience) sur la barre d'espace lorsque le phonème-cible était contenu dans le stimulus-cible et de ne rien faire dans le cas contraire (réponse de type Go/No Go). Il leur était également demandé de détecter le phonème-cible quelle que soit sa représentation orthographique. Par exemple, les items « gla**c**is » et « me**ss**ie » contiennent tous les deux le phonème /s/ alors que l'item « so**s**ie » correspond au phonème /z/. Cette précision était donnée à chaque participant, la détection

[32] Le point de fixation servait à focaliser l'attention du participant sur la zone d'apparition de l'image.

d'un phonème pouvant être modulée en fonction par la manière dont il est orthographié (Dijkstra, Roelofs, & Fieuws, 1995; voir Spinelli & Ferrand, 2005, pour une revue). Les participants avaient également pour tâche de prêter attention au signal acoustique ainsi qu'à l'information visuelle lorsque celle-ci était présente, l'attention portée à l'une ou l'autre modalité pouvant moduler l'intégration audiovisuelle des stimuli (e.g., Alsius, et al., 2005; Tiippana, et al., 2004). Le déroulement de chaque essai en fonction des conditions est représenté dans la Figure 24.

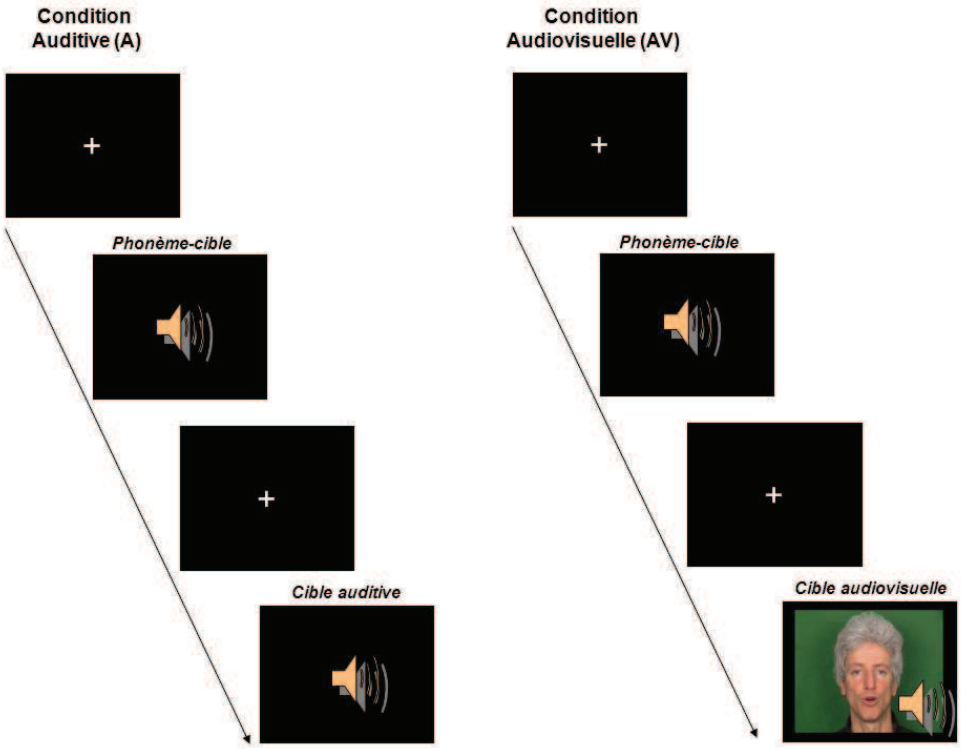


Figure 24. *Représentation schématique des différentes conditions expérimentales de l'Etude 1. Le mot ou le pseudo-mot cible pouvait être présenté à -9 dB, à -18 dB ou encore sans bruit (cf. post-test). Le mot ou le pseudo-mot cible pouvait soit contenir le phonème-cible (e.g., /pə/ dans /ʃapo/, « chapeau » ou /ʃapy/) soit ne pas le contenir (e.g., /pə/ pour /toʁty/, « tortue » vs. /toʁti/).En condition Audiovisuelle, le visage du locuteur en mouvement accompagnait le signal acoustique du mot ou du pseudo-mot cible.*

Pour chaque participant, 50 % des stimuli-cibles étaient présentés en condition AV (la moitié avec un RSB de -9 dB et l'autre moitié avec un RSB de -18 dB). Les 50 % des stimuli-cibles restant étaient présentés en condition AV (la moitié à -9 dB, le reste à -18 dB). L'expérience était divisée en deux blocs, l'un correspondant à la condition A, l'autre à la condition AV. L'ordre de ces blocs était contrebalancé entre les participants. Entre chaque bloc, un écran informait les participants du changement de modalité. A l'intérieur de chaque bloc, la première partie des stimuli était jouée avec un premier RSB (e.g., -9 dB ou -18 dB) et la seconde partie était présentée avec le second RSB. L'ordre de présentation des RSB était également contrebalancé entre les participants. Enfin, à l'intérieur de chaque condition,

l'ordre des stimuli était aléatoire. Une session d'entraînement comprenant 10 essais précédait le test à proprement parler. La génération des stimuli et la collecte du type et du temps de réponse était assurée par le logiciel E-Prime 2.0 (*Psychological Software Tools, Pittsburgh, PA*). La totalité de la passation durait 45 minutes environ.

3.2.2. Résultats

Le pourcentage de détections correctes et la moyenne des temps de réponse sur les détections correctes (mesurés à partir du début du phonème-cible) ont été calculés pour chaque participant et chaque paire d'items. Deux participants ont été retirés de l'analyse, ces derniers n'ayant donné aucune réponse dans la condition auditive à -18 dB. Une analyse de la variance (ANOVA) 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) x 2 (RSB : -9 dB vs. -18 dB) a été effectuée par participants (F_1) et par items (F_2), en échantillon appariés sur 79 participants, sur les temps de réponse d'une part et les détections correctes d'autre part. Les temps de réponse considérés comme aberrants (i.e., inférieur à 100 ms ou supérieur à 2500 ms) ou se situant à plus de 2 écart-types de la moyenne du participant pour chaque condition respective ont été exclus de l'analyse. Suite à cette opération 2.3 % des données totales ont été écartées.

3.2.2.1. Détections correctes

Les pourcentages de détections correctes pour chaque condition de l'Etude 1 sont présentés dans le Tableau 2.

Tableau 2. *Pourcentage de détections correctes en fonction des différentes conditions de l'Etude 1. L'erreur type est présentée entre parenthèses.*

RSB	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
-9 dB	A	68.2 (2.28)	69.4 (1.94)	-1.2
	AV	90.9 (1.18)	88.1 (1.33)	2.8
-18 dB	A	50.1 (2.16)	49.3 (2.03)	0.8
	AV	76.1 (2.14)	65.3 (2.05)	10.8

L'analyse statistique des données a tout d'abord permis de mettre en évidence un effet principal de la modalité de présentation, $F_1(1, 78) = 180.87$, $p < .001$, η^2_p [33] $= .70$; $F_2(1, 33) = 46.68$, $p < .001$, $\eta^2_p = .59$, révélant de meilleures performances pour la condition AV que pour la condition A. Un effet principal du facteur RSB a également été observé, $F_1(1, 78) = 199.34$, $p < .001$, $\eta^2_p = .72$; $F_2(1, 33) = 67.54$, $p < .001$, $\eta^2_p = .67$, révélant de meilleures performances à -9 dB qu'à -18 dB. Enfin, un effet principal du Statut Lexical significatif par participants, $F_1(1, 78) = 8.23$, $p < .005$, $\eta^2_p = .10$; et tendanciel par items, $F_2(1, 33) = 3.06$, $p = .08$, $\eta^2_p = .09$, a été obtenu, signifiant que les participants ont un pourcentage de détections correctes plus élevé lorsque le phonème-cible est contenu dans un mot que dans un pseudo-mot (i.e., effet de supériorité du mot).

Un effet d'interaction entre la Modalité et le Statut Lexical du stimulus-cible a été observé, $F_1(1, 78) = 10.34$, $p < .005$, $\eta^2_p = .12$; $F_2(1, 33) = 6.96$, $p < .05$, $\eta^2_p = .17$. Des comparaisons par paires ont permis de mettre en évidence un effet de supériorité du mot présent en condition AV, $F_1(1, 78) = 23.83$, $p < .001$, $\eta^2_p = .23$; $F_2(1, 33) = 10.21$, $p = .01$, $\eta^2_p = .24$, mais pas en condition A, F_1 et $F_2 < 1$. Un effet d'interaction entre le Statut Lexical et le RSB a également été obtenu, $F_1(1, 78) = 6.95$, $p = .01$, $\eta^2_p = .08$; $F_2(1, 33) = 4.69$, $p < .05$, $\eta^2_p = .12$. Dans la condition AV, des comparaisons par paires ont révélé que cet effet de supériorité du mot était clairement significatif à -18 dB, $F_1(1, 78) = 24.72$, $p < .001$, $\eta^2_p = .24$; $F_2(1, 33) = 11.48$, $p < .01$, $\eta^2_p = .27$ et à -9 dB mais seulement par participants, $F_1(1, 78) = 4.37$, $p < .05$, $\eta^2_p = .05$; $F_2(1, 33) = 2.54$, $p = .12$, $\eta^2_p = .07$. Aucun effet d'interaction entre la Modalité et le RSB n'a été obtenu, F_1 et $F_2 < 1$. Aucune interaction double entre ces trois facteurs n'a été observée, $F_1(1, 78) = 1{,}78$, $p = .19$, $\eta^2_p = .02$, $F_2(1, 33) = 2.54$, $p = .12$, $\eta^2_p = .03$.

Afin d'étudier si l'apport de l'information visuelle pouvait être modulée en fonction de la saillance dans le signal visuel des phonèmes à détecter, une analyse supplémentaire a été effectuée. Pour cela, le bénéfice apporté par la modalité visuelle a tout d'abord été calculé en soustrayant les moyennes des scores par items obtenus en modalité audiovisuelle à ceux obtenus en modalité auditive (tous les autres facteurs de l'analyse étant confondus). Des comparaisons post-hoc (test de Tuckey) ont permis de mettre en évidence que l'apport de l'information visuelle était plus important pour les phonèmes articulés à l'avant du conduit vocal (place d'articulation labiale, $M_{AV\text{-}A} = 31$ %) plutôt qu'à l'arrière (place d'articulation alvéodentale, $M_{AV\text{-}A} = 13$ %, $p < .005$).

[33] Le calcul de l'état-carré partiel (noté η^2_p) sera fourni en tant que mesure de la taille d'effet tout au long de ce manuscrit.

3.2.2.2. d'

Dans cette étude, la réponse demandée aux participants était de type Go/No Go. En conséquence, mesurer le pourcentage de détections correctes renseigne sur le nombre de fois où le participant a détecté correctement le phonème-cible et a fortiori sur le nombre de fois où il ne l'a pas détecté alors que celui-ci était présent dans le stimulus-cible (i.e., les *oublis*), mais ne nous renseigne pas sur ses capacités à rejeter correctement les essais où le phonème-cible n'était pas contenu dans le stimulus-cible (i.e., les *rejets corrects*) ni, a fortiori, sur le nombre de fois où il a perçu le phonème-cible alors que celui-ci était absent (i.e., les *fausses alarmes*). De ce fait, se limiter à analyser uniquement le pourcentage de détections correctes ne permet pas de se prémunir du fait que les participants aient pu adopter une certaine stratégie de réponse (e.g., prendre le risque de répondre plus fréquemment que la cible est présente plutôt qu'absente), augmentant/diminuant leur pourcentage de détections correctes mais également leur pourcentage de fausses alarmes.

Afin de prendre en compte le pourcentage de fausses alarmes dans nos analyses, un indice statistique, noté *d'*, a été calculé pour chaque participant[34]. Une ANOVA 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) x 2 (RSB : -9 dB vs. -18 dB) a été effectuée par participants (F_1) en échantillon appariés sur cet indice. Les *d'* moyennés pour chaque condition de l'Etude 1 sont présentés dans le Tableau 3.

Tableau 3. *Indice d' moyenné en fonction des différentes conditions de l'Etude 1. L'erreur type est présentée entre parenthèses.*

RSB	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
-9 dB	A	1.3 (0.09)	1.3 (0.08)	0
	AV	2.5 (0.06)	2.3 (0.06)	0.2
-18 dB	A	0.5 (0.09)	0.4 (0.08)	0.1
	AV	1.8 (0.08)	1.4 (0.09)	0.4

De la même manière que pour les détections correctes, l'analyse a permis de mettre en évidence de meilleures performances pour la condition AV, $F_1(1, 78) = 320.9$, $p < .001$, $\eta^2_p = .80$. Il a également été observé de meilleurs scores pour un RSB à -9 dB qu'à -18 dB, $F_1(1, 78) = 239.6$, $p < .001$, $\eta^2_p = .75$. Un effet de supériorité du mot a également été obtenu sur les *d'*, $F_1(1, 78) = 12.1$, $p < .001$, $\eta^2_p = .13$. Un effet d'interaction entre le Statut Lexical et

[34]La formule suivante a été utilisée : $d' = z(DC) - z(FA)$ où z représente l'inverse de la distribution de la loi normale centrée réduite, *DC* et *FA* font respectivement référence à la probabilité moyenne de Détections Correctes et de Fausses Alarmes.

la Modalité a aussi été observé, $F_1(1, 78) = 4.9$, $p < .05$, $\eta^2_p = .06$. Des comparaisons par paires ont permis de mettre en évidence un effet de supériorité du mot plus important en AV, $F_1(1, 78) = 20.13$, $p < .001$, $\eta^2_p = .20$, qu'en A, $F_1 < 1$. Un effet d'interaction significatif entre le Statut Lexical et le RSB a aussi été obtenu, $F_1(1,78) = 4.5$, $p < .05$, $\eta^2_p = .05$, mettant en évidence un effet de supériorité du mot en condition AV (à -9 dB, $F_1(1,78) = 7.21$, $p < .005$, $\eta^2_p = .08$, et à -18 dB, $F_1(1,78) = 16.72$, $p < .001$, $\eta^2_p = .15$) et pas en condition A (à -9 dB, $F_1 < 1$, et à -18 dB, $F_1(1,78) = 1.69$, $p = .20$, $\eta^2_p = .02$).

3.2.2.3. Temps de réponse

Les temps de réponse moyens pour chaque condition de l'Etude 1 sont présentés dans le Tableau 4.

Tableau 4. *Temps de réponse moyens (en millisecondes, ms) en fonction des différentes conditions de l'Etude 1. L'erreur type est présentée entre parenthèses.*

RSB	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
-9 dB	A	808 (31.4)	809 (32.7)	-4
	AV	677 (27.9)	688 (22.1)	-11
-18 dB	A	886 (38.1)	910 (41.5)	-24
	AV	748 (27.8)	761 (33.4)	-13

Un effet principal de la Modalité de Présentation a été obtenu $F_1(1, 78) = 32.27$, $p < .001$, $\eta^2_p = .29$; $F_2(1, 33) = 62.32$, $p < .001$, $\eta^2_p = .65$, révélant que les participants étaient plus rapides pour détecter le phonème-cible pour la condition AV. Un effet principal du facteur RSB a également été observé, $F_1(1, 78) = 31.01$, $p < .001$, $\eta^2_p = .28$; $F_2(1, 33) = 8.51$, $p < .01$, $\eta^2_p = .20$, signifiant que les participants étaient plus rapides pour effectuer la tâche à -9 dB qu'à -18 dB. Cependant, aucun effet principal du Statut Lexical $F_1(1, 78) = 1.02$, $p = .27$, $\eta^2_p = .02$; $F_2(1, 33) = 1.44$, $p = .23$, $\eta^2_p = .04$, ni d'interaction (simple ou double) entre les facteurs cités précédemment n'a été obtenu, (tous les $F_1 < 1$).

Afin d'étudier si l'apport de l'information visuelle pouvait être modulée en fonction de la saillance dans le signal visuel des phonèmes à détecter, une analyse supplémentaire a été effectuée. Pour cela, le bénéfice possible apporté par la modalité visuelle a été calculé en soustrayant les moyennes des temps de réponse par items obtenus en modalité auditive à ceux obtenus en modalité audiovisuelle (tous les autres facteurs de l'analyse étant

confondus). Des comparaisons post-hoc (test de Tuckey) ont permis de mettre en évidence que les participants étaient d'autant plus rapides pour détecter un phonème en modalité audiovisuelle, par rapport à la modalité auditive seule, si les phonèmes étaient articulés à l'avant du conduit vocal (place d'articulation labiale, $M_{\text{A-AV}} = 172$ ms) plutôt qu'à l'arrière (place d'articulation alvéodentale, $M_{\text{A-AV}} = 101$ ms, $p < .05$).

3.2.3. Discussion

L'objectif de l'Etude 1 était d'explorer le rôle de la gestualité oro-faciale dans le processus d'accès au lexique, en présence d'une information auditive congruente. Pour cela, nous avons utilisé une tâche de détection de phonèmes consonantiques à effectuer dans des mots ou des pseudo-mots, présentés en modalité AV ou A, avec différents niveaux de bruits ans le signal acoustique (-9 dB vs. -18dB). Rappelons que différents niveaux de bruits ont donc été utilisés afin de mettre en évidence l'apport de de l'information visuelle en présence de l'information auditive.

Les résultats montrent que les participants avaient de meilleures performances (i.e., des pourcentages de détections correctes plus élevés et des temps de réponse plus courts) dans la condition AV que dans la condition A. Un avantage a aussi été observé lorsque le stimuli-cible était présenté à -9 dB plutôt qu'à -18 dB, sur l'ensemble de ces mesures. Les résultats mettent également en évidence des pourcentages de détections correctes et des *d'* plus élevés lorsque les phonèmes-cibles étaient présentés dans un mot plutôt que dans un pseudo-mot. Conformément à nos prédictions, cet effet de supériorité du mot était d'autant plus important lorsque l'information visuelle était présente (i.e., dans la condition AV par rapport à la condition A). Ces données suggèrent donc que l'information visuelle contribue au processus d'accès au lexique, en présence d'une information auditive congruente. Il faut cependant remarquer que ce résultat n'a pas été observé sur les temps de réponse, comme cela avait été supposé. De plus, aucun effet principal de supériorité du mot n'a pu être mis en évidence sur cette mesure, alors qu'il est observé dans d'autres études (e.g., Frauenfelder, et al., 1990). En conséquence, afin de s'assurer que cette absence d'effet lexical ne soit pas due à une spécificité du matériel utilisé (stimuli, paradigme, etc.), l'Etude 1 a été reconduite avec d'autres participants, dans des conditions de perception du signal acoustique similaires telles que proposées dans ces études, c'est-à-dire en l'absence de détérioration du signal acoustique.

3.2.4. Post-test

3.2.4.1. Participants

Trente-sept participants (dont 28 femmes et 9 hommes) âgés de 18 à 32 ans (M = 21.8 ans) ont été recrutés pour cette étude. Tous étaient de langue maternelle française. Aucun d'entre eux n'avait de déficit sur le plan auditif et tous avaient une vision normale ou corrigée. La majorité d'entre eux étaient étudiants en Psychologie à l'Université Pierre Mendès France de Grenoble et recevaient un bon d'expérimentation en échange de leur participation. Aucun d'entre eux n'avait participé à l'étude précédente.

3.2.4.2. Stimuli et procédure

Les mêmes stimuli que dans l'expérience précédente ont été utilisés pour construire ce post-test, excepté que le RSB de ces derniers n'était pas modifiés par rapport à l'enregistrement original (i.e., condition sans bruit). Du fait de l'absence de manipulation du facteur RSB, l'ensemble des items était distribué selon 2 listes, correspondant aux 2 conditions de présentation des stimuli (A vs. AV). Chaque liste contenait 17 paires d'items expérimentaux et 20 paires d'items de remplissage. La procédure et les contrebalancements employés étaient identiques à ceux utilisés précédemment. Les stimuli étaient générés par le logiciel E-Prime 2.0 (*Psychological Software Tools, Pittsburgh, PA*).

3.2.5. Résultats

Le pourcentage de détections correctes et la moyenne des temps de réponse (mesurés à partir du début du phonème-cible) sur les détections correctes ont été calculés pour chaque participant et chaque paire d'items. Une ANOVA 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) a été effectuée sur les temps de réponse d'une part et les détections correctes, par participants (F_1) et par items (F_2), en échantillon appariés sur les 37 participants. Les temps de réponse considérés comme aberrants (i.e., inférieur à 100 ms ou supérieur à 2500 ms) ou se situant à plus de 2 écart-types de la moyenne du participant au sein de chaque condition ont été exclus de l'analyse. Un participant dont l'ensemble des performances était situé à plus (temps de réponse) ou moins (réponses correctes) de 2 écart-types de la moyenne a été exclu de l'analyse. Suite à cette opération 1 % des données totales *a* été écarté. Les pourcentages de détections correctes et les temps de réponse pour chaque condition du post-test sont présentés dans le Tableau 5.

Tableau 5. *Pourcentage de détections correctes (DC) et temps de réponse moyens (TR, en ms) en fonction des différentes conditions du post-test de l'Etude 1. L'erreur type est présentée entre parenthèses.*

Variable dépendante	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
DC	A	95.1 (1.17)	96.1 (1.22)	-1
	AV	95.9 (1.15)	94.4 (1.37)	1.5
TR	A	498 (10.7)	509 (7.9)	12
	AV	471 (12.2)	490 (12.4)	19

Le pourcentage de détections correctes étant proche de 100 % pour l'ensemble des conditions ($M = 94.8$ %) l'analyse statistique n'a révélé aucun effet significatif sur cette mesure[35] (tous les $F_1 < 1$). L'analyse sur les temps de réponse a mis en évidence un effet du statut lexical des items, $F_1(1, 35) = 7.18$, $p < .05$, $\eta^2_p = .17$; $F_2(1, 33) = 4.73$, $p < .05$, $\eta^2_p = .13$. A l'instar des études utilisant la tâche de détection de phonèmes en modalité auditive (e.g., Frauenfelder et al., 1990), les participants étaient plus rapides pour détecter le phonème-cible lorsque le stimulus-cible était un mot. Aucun effet principal de la Modalité de Présentation ou d'interaction entre les 2 facteurs de l'analyse n'a pu être mis en évidence, tous les $F_1 < 1$.

Afin d'étudier si l'apport de l'information visuelle pouvait être modulée en fonction de la saillance dans le signal visuel des phonèmes à détecter, une analyse supplémentaire a été effectuée sur les pourcentages de réponses correctes et les temps de réponse. Le bénéfice apporté par la modalité visuelle a été calculé sur ces deux mesures de la même manière qu'en condition de détérioration du signal acoustique. Les comparaisons post-hoc n'ont révélé aucun bénéfice significatif de la présence de la modalité visuelle pour la détection des phonèmes labiaux par rapport à ceux présentant une place d'articulation alvéodentale (tous les $p > .05$).

[35] Un indice *d'* a été également été calculé puis soumis à une ANOVA 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot). De la même manière que sur les pourcentages de détections correctes, aucun effet principal ni d'interaction entre les 2 facteurs de l'analyse ne s'est révélé significatif, tous les $F < 1$.

3.2.6. Discussion

L'objectif de l'Etude 1 était d'explorer l'apport de l'information visuelle (i.e., mouvements articulatoires de la face issus du geste de parole) à la perception de l'information auditive (i.e., signal acoustique issus du geste de parole) dans le processus de reconnaissance de mots. Afin d'étudier cette question, une tâche de détection de phonèmes consonantiques a été utilisée. Dans ce paradigme, le Statut Lexical (mot vs. pseudo-mot), la Modalité de Présentation (AV vs. A) mais également le niveau de détérioration du signal acoustique (facteur RSB : -9 dB vs.-18 dB ; post-test en l'absence de bruit) des stimuli-cibles ont été manipulés.

3.2.6.1. Apport de l'information visuelle

Premièrement, les résultats obtenus dans cette étude sont en accord avec ceux observés par Benoît et al. (1994). En effet, ces données montrent que lorsque l'information acoustique était artificiellement détériorée par du bruit blanc (i.e., à -9 dB et à -18 dB), les participants présentaient un pourcentage de détections correctes plus élevé lorsque les stimuli-cibles étaient présentés en condition AV plutôt qu'en condition A. Cet avantage de la modalité AV par rapport à la modalité A (lorsque le signal acoustique est dégradé) a largement été étudié et répliqué dans la littérature. Remarquons ici que les pourcentages de détections correctes obtenus en condition A à -9 dB étaient similaires à ceux obtenus en condition AV à -18 dB (.M =68.9 % vs. M =70.7 %, respectivement). Il semblerait donc que dans cette étude, le gain d'intelligibilité observé par la présence de l'information visuelle soit environ de 9 dB, soit de l'ordre de ceux observés dans la littérature (e.g., M = 11 dB, MacLeod & Summerfield, 1987).

Ce bénéfice a notamment été expliqué par le fait que lorsque l'information auditive est détériorée, le signal visuel devient alors complémentaire (en termes de type d'information véhiculée) du signal acoustique (Summerfield, 1987). Cela signifie que l'information auditive qui est facilement masquée par du bruit (e.g., la place d'articulation, e.g., Miller & Nicely, 1955) est généralement saillante sur le plan articulatoire et peut, en conséquence, être aisément restaurée lorsque les mouvements des articulateurs tels que les lèvres, les dents, la langue et la mâchoire sont visibles (Robert-Ribes, et al., 1998), (voir la section 1.2.3 pour plus de détails à ce sujet). En accord avec cette explication, les comparaisons post-hoc ont révélé que dans notre étude, l'apport de l'information visuelle était plus important pour la détection des phonèmes les plus saillants sur le plan articulatoire, c'est-à-dire articulés à l'avant du conduit vocal. Les résultats obtenus sur les temps de réponse montrent que le

processus de détection de phonèmes était plus rapidement effectué en condition AV, majoritairement pour les phonèmes les plus visibles, mais ce bénéfice n'a été observé qu'en présence de bruit dans le signal acoustique. Ces données mettent en évidence que la disponibilité des informations articulatoires permet d'augmenter l'intelligibilité des phonèmes en améliorant non seulement leur identification (e.g., Benoît, et al., 1994) mais également en *accélérant* leur processus de détection. Cet avantage peut être expliqué de deux manières. Une première hypothèse serait que l'information visuelle permet de renforcer l'activation des unités de décision phonémiques, permettant de faciliter (e.g., Benoît, et al., 1994) mais également d'*accélérer* le processus de détection de phonème (voir Cox, et al., 1999, pour des résultats similaires). Une autre explication possible de ce phénomène serait de supposer que l'information visuelle accélère le processus de détection de phonèmes car celle-ci permet d'anticiper sa conséquence acoustique (e.g., Cathiard, 1994; Jesse & Massaro, 2010; Munhall & Tohkura, 1998; Smeele, 1994). Les résultats présents ne nous permettent cependant pas d'argumenter en faveur de l'une ou l'autre hypothèse.

3.2.6.2. Apport de l'information lexicale

Les résultats observés permettent également de répliquer l'effet de supériorité du mot généralement retrouvé dans la littérature consacrée à l'étude de l'accès au lexique (e.g., Frauenfelder, et al., 1990). En effet, dans notre étude, les phonèmes-cibles consonantiques étaient mieux reconnus lorsque ceux-ci étaient présentés dans le contexte d'un mot plutôt que dans celui d'un pseudo-mot. En présence de bruit dans le signal acoustique, cet effet s'est révélé être significatif en condition AV sur le pourcentage de réponses correctes mais pas dans la condition A. En l'absence de bruit, les résultats issus du post-test ont permis de mettre en avant un effet principal du statut lexical, de manière indifférenciée pour la condition A et AV sur les temps de réponse (cf. section 3.2.6).

L'effet de supériorité du mot est généralement interprété comme le signe d'une influence de l'information lexicale sur le processus de traitement des phonèmes par (1) un retour d'activation du niveau lexical vers les représentations phonologiques (processus de « feedback » approche « top-down » postulée par exemple par le modèle TRACE, McClelland & Elman, 1986) ou par (2) une diffusion de l'activation du niveau lexical vers le niveau de décision phonémique (approche « bottom-up » postulée par exemple par le modèle Merge, Norris, et al., 2000). Nous tenons à souligner que l'objectif de cette étude n'était pas d'essayer de déterminer le sens (bottom-up ou top-down) de l'influence du contexte lexical et

que les résultats obtenus dans cette étude sont compatibles avec chacune de ces interprétations.

3.2.6.3. Apport combiné de l'information visuelle et lexicale

En situation bruitée (i.e., à -9 dB et surtout à -18 dB), les résultats montrent que l'effet de supériorité du mot était uniquement présent dans la condition AV. Ce résultat suggère que l'information visuelle facilite non seulement le processus de détection de phonèmes mais *participe* également à l'accès lexical, lorsque le signal acoustique de parole est détérioré. En effet, l'effet lexical observé dans cette étude ne peut être uniquement expliqué par la présence de l'information auditive, puisqu'aucune influence du statut lexical n'a été observée en condition A, pour les deux conditions de bruit.

Lorsque l'information auditive était intacte (post-test), aucune interaction entre le statut lexical et la modalité de présentation des stimuli-cibles n'a cependant pu être mis en évidence. Cette absence d'interaction sur les pourcentages de réponses correctes est probablement due à un effet plafond des performances des participants, en modalité A comme en modalité AV ($M = 95.1$ % vs. $M = 94.5$ %, respectivement). L'absence d'interaction sur les temps de réponses suggère que lorsque les conditions de perception de la parole sont optimales (i.e., lorsque le signal acoustique est intact), l'information auditive est suffisante pour accéder efficacement et suffisamment rapidement aux représentations lexicales (voir Spinelli & Ferrand, 2005, pour une revue récente des travaux sur la reconnaissance auditive des mots parlés). Ainsi, il serait envisageable que l'information extraite des mouvements articulatoires du visage de notre interlocuteur pourrait participer à l'activation des représentations lexicales principalement lorsque la reconnaissance du signal de parole est rendue difficile (e.g., lorsque l'information auditive est détériorée).

3.2.7. Conclusions

Dans cette étude, en présence de bruit, nous pouvons remarquer une absence d'effet lexical et d'interaction entre le statut lexical et la modalité de présentation sur les temps de réponse. Egalement, aucun effet lexical (sur les temps de réponse comme sur les pourcentages de réponses correctes) n'a pu être mis en évidence dans l'Etude 1 pour la condition A dans les 2 conditions de bruit (i.e., à -9 dB et -18 dB). Les résultats du post-test ont mis en évidence un effet lexical sur les temps de réponse en modalité A comme en modalité AV, écartant l'hypothèse que l'absence de cet effet soit due à une spécificité du matériel et/ou de la procédure utilisée. L'hypothèse d'une difficulté générale à effectuer la

tâche dans le bruit (M = 69.7 %, le hasard étant situé à 50 % dans cette étude) pourrait expliquer qu'aucun effet lexical n'a pu être observé sur les temps de réponse. Cette dernière permettrait également d'expliquer l'absence d'effet de supériorité du mot pour la condition A à -18 dB, les performances des participants dans cette condition n'étaient pas différentes du hasard (M = 50.1 % pour les mots vs. M = 49.3 % pour les pseudo-mots, tous les $t < 1$). Cette explication peut plus difficilement justifier ce phénomène pour la condition A à -9 dB. En effet, dans cette condition, le pourcentage de détection correctes (M = 68.2 % pour les mots vs. M = 69.4 % pour les pseudo-mots) est comparable à celui obtenu en condition AV à -18 dB, où un effet de supériorité du mot significatif est observé (M = 76.1 % pour les mots vs. M = 65.3 % pour les pseudo-mots). Nous n'avons aucune explication à ce jour permettant de justifier de cette absence d'effet en modalité A à -9 dB.

3.2.8. Objectifs de l'Etude 2

Afin de tester l'hypothèse d'une difficulté générale de la tâche, cette même étude a été effectuée une nouvelle fois, en utilisant des phonèmes-cibles vocaliques plutôt que consonantiques (Etude 2). En effet, les voyelles semblent être des entités plus facilement détectables dans le signal que les consonnes (Ladefoged, 2001). De plus, leur reconnaissance semble mieux résister à l'ajout de bruit dans le signal acoustique (Nooteboom & Doodeman, 1984, cité par Cutler, Sebastián-Gallés, Soler-Vilageliu, & van Ooijen, 2000). Nous avons donc décidé de mener une seconde étude (Etude 2) afin de de tester si le fait de proposer des phonèmes-cibles vocaliques tout en gardant le même niveau de bruit que précédemment nous permet d'obtenir les effets attendus sur les temps de réponse, en modalité audiovisuelle tout du moins. Pour cela, nous avons utilisé une tâche de détection de phonèmes vocaliques dans des mots et des pseudo-mots, en modalité A ou audiovisuelle AV en l'absence (sans bruit) ou en présence de bruit blanc dans le signal acoustique (à -9 dB, et à -18 dB). Conformément à nos précédentes hypothèses, nous nous attendions tout d'abord à répliquer les effets obtenus pour l'Etude 1, c'est-à-dire à obtenir un effet de supériorité du mot plus important pour la condition AV que pour la condition A, lorsque l'information auditive est détériorée. En utilisant des phonèmes-cibles vocaliques plutôt que consonantiques, nous nous attendons à observer cet effet d'interaction également sur les temps de réponse. Pour cette même raison, nous devrions aussi obtenir un effet de supériorité du mot en condition A lorsque l'information acoustique est détériorée.

3.3. ETUDE 2 : INFLUENCE DE L'INFORMATION VISUELLE ET LEXICALE DANS LE PROCESSUS DE DETECTION DE PHONEMES VOCALIQUES

3.3.1. Méthode

3.3.1.1. Participants

Soixante participants (dont 43 femmes et 17 hommes) âgés de 18 à 38 ans (M = 22 ans) ont été recrutés pour cette étude. Tous étaient de langue maternelle française. Aucun d'entre eux n'avait de déficit sur le plan auditif et tous avaient une vision normale ou corrigée. La moitié d'entre eux étaient étudiants en Psychologie à l'Université Pierre Mendès France de Grenoble et recevaient un bon d'expérimentation en échange de leur participation. Aucun d'entre eux n'avait participé à l'Etude 1.

3.3.1.2. Stimuli

3.3.1.2.1. Items expérimentaux

Un corpus de 90 paires de mots/pseudo-mots bisyllabiques de type CVCV a été sélectionné en utilisant la base de données LEXIQUE 2 (New, et al., 2001, cf. Annexe B). La fréquence moyenne des mots porteurs du phonème-cible était de 62.03 opm, (LEXIQUE 2, New, et al., 2001). Chacun de ces items comportait un des cinq phonème-cibles consonantiques sélectionnés pour cette étude : ceux-ci pouvaient engendrer soit un mouvement d'arrondissement (/o/, /u/, /y/) soit un mouvement d'étirement des lèvres (/i/, /e/). Le phonème-cible était toujours situé à la fin de la seconde syllabe (e.g., le phonème-cible /**o**/ dans /bat**o**/, « bat**eau** » ou /nat**o**/), afin de s'assurer que chaque phonème-cible vocalique à détecter soit, à l'intérieur d'une même paire, inséré dans un contexte consonantique similaire. Comme dans l'Etude 1, ce dernier était toujours situé à la fin du mot (Frauenfelder, et al., 1990). La durée moyenne des mots ne différait pas de celle des pseudo-mots (M_{mots} = 834 ms ; $M_{pseudo\text{-}mots}$ = 888 ms, $t(178) = 1.9$, $p > .05$).

3.3.1.2.2. Items de remplissage

Afin que les participants ne répondent pas que le phonème-cible est présent à chaque essai, 45 paires de mot/pseudo-mot, correspondant aux items de remplissage (i.e., ne contenant pas le phonème-cible à détecter), ont été construits en utilisant les mêmes phonèmes que les items expérimentaux. Cependant, à la différence de ces derniers, les items de remplissage

étaient toujours associés à un phonème différent ne correspondant à aucun de ceux contenus dans l'item en question (e.g., le phonème-cible /o/, pour /mɛʁsi/, « merci » vs. /lɛʁsi/). Ces phonèmes de remplissage étaient choisis afin de différer sur le plan acoustique et articulatoire de l'ensemble des phonèmes contenus dans l'item cible. La fréquence moyenne de ces items était de 32.26 opm.

3.3.1.2.3. Enregistrement des stimuli

L'ensemble des stimuli (mots, pseudo-mots et phonèmes-cibles) a été enregistré et numérisé avec le même matériel et les mêmes procédures que dans l'Etude 1. Ils étaient prononcés par une locutrice entraînée, de langue maternelle française. Elle avait pour consigne d'initier la production de chaque item en partant avec la bouche fermée et de ne pas cligner des yeux durant la prononciation de chacun d'eux. Elle était placée devant un fond vert et seule sa tête (des sourcils au bas du menton) était visible (cf. Figure 25).

3.3.1.3. Procédure

La procédure et que le lieu de passation de l'expérience étaient identiques à l'Etude 1. Contrairement à l'Etude 1, la présentation de l'item-cible en modalité A était accompagnée de l'image fixe du visage la locutrice, bouche fermée. Cette modification a été apportée afin de s'assurer que le simple fait de voir le visage de la locutrice en l'absence de mouvement articulatoire n'explique pas les éventuelles différences de performances entre la modalité AV et A. L'image fixe comme la composante vidéo des stimuli étaient toujours présentées en noir et blanc. Le déroulement de chacun des essais en fonction des conditions est représenté dans la Figure 25.

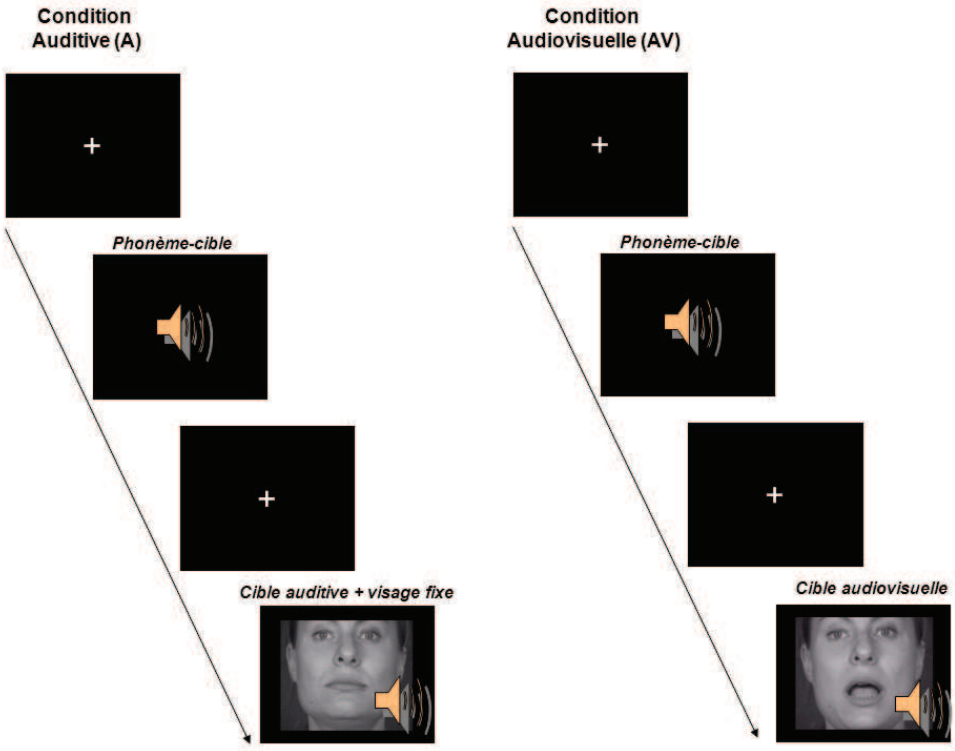


Figure 25. *Représentation schématique des différentes conditions expérimentales de l'Etude 2. Le mot ou le pseudo-mot cible pouvait être présenté à -9 dB, à -18 dB ou encore sans bruit. Le mot ou le pseudo-mot cible pouvait soit contenir le phonème-cible (e.g., /o/ dans /bato/, « bateau » ou /nato/) soit ne pas le contenir (e.g., /o/ pour /mɛʁsi/, « merci » vs. /lɛʁsi/).En condition A, la présentation du mot ou du pseudo-mot cible était accompagnée du visage fixe de la locutrice.*

A la différence de l'Etude 1, chaque participant passait les trois conditions de RSB différentes : sans bruit, -9 dB et -18 dB. En conséquence, chaque paire de mot/pseudo-mot était distribuée selon 6 listes correspondant aux 6 conditions de présentation des stimuli : A sans bruit ; A à -9 dB ; A à -18 dB ; AV sans bruit ; AV à -9 dB ; AV à -18 dB. Chaque liste contenait 15 paires d'items expérimentaux et 15 paires d'items de remplissage. Chaque item (expérimental ou de remplissage) était présenté une seule fois à chaque participant. Pour chaque participant, 50 % des stimuli-cibles étaient présentés en condition AV (un tiers sans bruit, un tiers à -9 dB, le reste à -18 dB) et 50 % en condition A (un tiers sans bruit, un tiers à -9 dB, le reste à -18 dB). Une session d'entraînement comprenant 8 essais précédait la phase test à proprement parler. La génération des stimuli et la collecte du type et du temps de réponse était assurée par le logiciel E-Prime 2.0 (*Psychological Software Tools, Pittsburgh, PA*). La totalité de la passation durait 40 minutes environ.

3.3.2. Résultats

Le pourcentage de détections correctes et la moyenne des temps de réponse (mesurés à partir du début du phonème-cible) sur les détections correctes ont été calculés pour chaque participant et chaque paire d'items. Afin de respecter l'homogénéité des variances, deux analyses séparées ont été effectuées sur la condition sans bruit d'une part et les conditions « en présence de bruit » (i.e., à -9 dB et -18 dB) d'autre part.

3.3.2.1. En situation bruitée

Une analyse de la variance (ANOVA) 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) x 2 (RSB : -9 dB vs. -18 dB) a donc été effectuée, par participants (F_1) et par items (F_2), en échantillon appariés sur 60 participants, sur les temps de réponse d'une part et les détections correctes d'autre part. Les temps de réponse considérés comme aberrants (i.e., inférieur à 100 ms ou supérieur à 2500 ms) ou se situant à plus de 2 écart-types de la moyenne du participant au sein de chaque condition ont été exclus de l'analyse. Ainsi, 3.2 % des données totales ont été écartés.

3.3.2.1.1. Détections correctes

Les pourcentages de détections correctes pour les conditions bruitées de l'Etude 2 sont présentés dans le Tableau 6.

Tableau 6. *Pourcentage de détections correctes pour chaque condition bruitée de l'Etude 2. L'erreur type est présentée entre parenthèses.*

RSB	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
-9 dB	A	78.6 (1.5)	68.1 (1.7)	10.5
	AV	93.7 (0.9)	91.7 (1.1)	3
-18 dB	A	55.9 (2.3)	56.7 (2.2)	-0.8
	AV	90.4 (1.2)	86.8 (1.5)	3.6

L'analyse statistique des données a permis de mettre en évidence un effet principal de la modalité de présentation, $F_1(1, 59) = 392.3$, $p < .001$, $\eta^2_p = .87$; $F_2(1, 89) = 340.3$, $p < .001$, $\eta^2_p = .79$, révélant de meilleures performances pour la condition AV que pour la condition A. Un effet principal du facteur RSB a également été observé, $F_1(1, 59) = 98.4$, $p < .001$, $\eta^2_p = .63$; $F_2(1, 89) = 57.5$, $p < .001$, $\eta^2_p = .39$, révélant de meilleures performances à -9 dB qu'à

-18 dB. Enfin, un effet principal du Statut Lexical a été obtenu, $F_1(1, 59) = 20$, $p < .001$, $\eta^2_p = .25$; $F_2(1, 89) = 8.74$, $p < .005$, $\eta^2_p = .09$, signifiant que les participants ont un pourcentage de détections correctes plus élevé lorsque le phonème-cible est contenu dans un mot (i.e., effet de supériorité du mot).

Un effet d'interaction double entre la Modalité, le Statut Lexical et le Rapport Signal sur Bruit du stimulus-cible a été observé, $F_1(1, 59) = 12$, $p < .001$, $\eta^2_p = .17$; $F_2(1, 89) = 13.4$, $p < .001$, $\eta^2_p = .13$. Des comparaisons par paires ont permis de mettre en évidence un effet de supériorité du mot plus important en A qu'en AV à -9 dB, $F_1(1, 59) = 13.9$, $p < .001$, $\eta^2_p = .19$; $F_2(1, 89) = 9.6$, $p < .005$, $\eta^2_p = .10$. A -18 dB, les résultats suggèrent à l'inverse un effet lexical plus important en AV qu'en A, $F_1(1, 59) = 3.08$, $p = .08$, $\eta^2_p = .05$; $F_2(1, 89) = 2.51$, $p = .11$, $\eta^2_p = .03$, mettant en évidence un effet lexical significatif à -18 dB en AV, $F_1(1, 59) = 6.34$, $p < .05$, $\eta^2_p = .10$; $F_2(1, 89) = 3.95$, $p = .05$, $\eta^2_p = .04$, mais pas en A, $F_1(1, 59) < 1$.

3.3.2.1.2. *d'*

Afin de prendre en compte le pourcentage de fausses alarmes dans nos analyses, un *d'* a été calculé pour chaque participants. Une ANOVA 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) x 2 (RSB : -9 dB vs. -18 dB) a donc été effectuée en échantillon appariés par participants (F_1) sur cet indice. Les *d'* moyennés pour les conditions bruitées de l'Etude 2 sont présentés dans le Tableau 7.

Tableau 7. *Indice d' moyenné pour chaque condition bruitée de l'Etude 2. L'erreur type est présentée entre parenthèses.*

RSB	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
-9 dB	A	1.74 (0.07)	1.24 (0.07)	0.5
	AV	3.04 (0.06)	2.84 (0.07)	0.2
-18 dB	A	0.81 (0.07)	0.72 (0.05)	0.11
	AV	2.69 (0.06)	2.42 (0.07)	0.27

De la même manière que sur les détections correctes, l'analyse a permis de mettre en évidence de meilleures performances pour la modalité AV, $F_1(1, 59) = 946.5$, $p < .001$, $\eta^2_p = .94$. Il a également été observé de meilleurs scores à -9 dB qu'à -18 dB, $F_1(1, 59) = 163.2$, $p < .001$, $\eta^2_p = .73$. Un effet de supériorité du mot a également été obtenu sur les *d'*, $F_1(1, 59) = 38.24$, $p < .001$, $\eta^2_p = .39$. Un effet d'interaction double entre la Modalité, le Statut

Lexical et le Rapport Signal sur Bruit du stimulus-cible a été observé, $F_1(1, 59) = 7.3$, $p < .005$, $\eta^2_p = .11$. Des comparaisons par paires ont permis de mettre en évidence un effet de supériorité du mot plus important en A qu'en AV à -9 dB, $F_1(1, 59) = 5.99$, $p < .05$, $\eta^2_p = .09$. A l'inverse, à -18 dB, les résultats mettent en un effet lexical en AV, $F_1(1, 59) = 5.04$, $p < .05$, $\eta^2_p = .08$; mais pas en A, $F_1(1, 59) < 1$. Notons que les résultats obtenus sur les *d'* étant identiques à ceux obtenus sur les réponses correctes, ces résultats ne seront pas spécifiquement commentés dans la suite de cette étude.

3.3.2.1.3. Temps de réponse

Les temps de réponse moyens pour chaque condition de l'Etude 2 sont présentés dans le Tableau 8. Afin de respecter l'homogénéité des variances, une transformation logarithmique des temps de réponse moyens a été effectuée avant l'analyse.

Tableau 8. *Temps de réponse moyens (en millisecondes, ms) pour chaque condition bruitée de l'Etude 2. L'erreur type est présentée entre parenthèses.*

RSB	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
-9 dB	A	482 (14)	513 (16)	31
	AV	423 (12)	471 (11)	48
-18 dB	A	585 (20)	577 (16)	-8
	AV	481 (12)	508 (14)	27

Un effet principal de la Modalité de Présentation a été obtenu $F_1(1, 59) = 57.1$, $p < .001$, $\eta^2_p = .49$; $F_2(1, 89) = 125.6$, $p < .001$, $\eta^2_p = .59$, révélant que les participants étaient plus rapides pour détecter le phonème-cible pour la condition AV. Un effet principal du facteur RSB a également été observé, $F_1(1, 59) = 91.03$, $p < .001$, $\eta^2_p = .60$; $F_2(1, 89) = 91.03$, $p < .001$, $\eta^2_p = .61$, signifiant que les participants étaient plus rapides pour effectuer la tâche à -9 dB qu'à -18 dB. Un effet principal du Statut Lexical a également été obtenu, $F_1(1, 59) = 44.02$, $p < .001$, $\eta^2_p = .43$; $F_2(1, 89) = 19.5$, $p < .001$, $\eta^2_p = .18$. Un effet d'interaction a été obtenu entre le statut lexical et la modalité de présentation, $F_1(1, 59) = 10.81$, $p < .005$, $\eta^2_p = .15$; $F_2(1, 89) = 11.3$, $p < .005$, $\eta^2_p = .11$, mettant en évidence un effet de supériorité du mot plus important en modalité AV, $F_1(1, 59) = 68.7$, $p < .001$, $\eta^2_p = .54$; $F_2(1, 89) = 41.6$, $p < .001$, $\eta^2_p = .32$, qu'en modalité A seule, $F_1(1, 59) = 4.01$, $p < .05$, $\eta^2_p = .06$; $F_2(1, 89) = 2.1$, $p = .15$, $\eta^2_p = .02$.

3.3.2.2. En l'absence de détérioration du signal acoustique

Une analyse de la variance (ANOVA) 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) a donc été effectuée par participants (F_1) et par items (F_2), en échantillon appariés sur 60 participants, sur les temps de réponse d'une part et les détections correctes d'autre part. Les temps de réponse considérés comme aberrants (i.e., inférieur à 100 ms ou supérieur à 2500 ms) ou se situant à plus de 2 écart-types de la moyenne du participant au sein de chaque condition ont été exclus de l'analyse. Ainsi, 0.9 % des données totales a été écarté.

Les pourcentages de détections correctes ainsi que les temps de réponse pour chaque condition non bruitée de l'Etude 2 sont présentés dans le Tableau 9.

Tableau 9. *Pourcentage de détections correctes (DC) et temps de réponse moyens (TR, en ms) pour les conditions non bruitées de l'Etude 2. L'erreur type est présentée entre parenthèses.*

Variable dépendante	Modalité	Mot	Pseudo-mot	Effet de supériorité du mot
DC	A	98.8 (0.3)	97.4 (0.5)	1.5
	AV	100 (0)	99.2 (0.3)	0.8
TR	A	354 (9.6)	411 (13.6)	57
	AV	333 (9.8)	389 (9.4)	56

Le pourcentage de détections correctes étant proche de 100 % pour l'ensemble des conditions ($M = 98.9$ %) aucune analyse statistique n'a été effectuée sur cette mesure[36]. L'analyse sur les temps de réponse a mis en évidence un effet du statut lexical des items, $F_1(1, 59) = 7.88$, $p < .01$, $\eta^2_p = .12$; $F_2(1, 89) = 21.3$, $p < .001$, $\eta^2_p = .19$, répliquant l'effet de supériorité du mot observé dans la littérature en modalité auditive (e.g., Frauenfelder et al., 1990) et dans l'Etude 1. Un effet principal de la Modalité de Présentation a été observé, $F_1(1, 59) = 106.5$, $p < .001$, $\eta^2_p = .64$; $F_2(1, 89) = 85.1$, $p < .001$, $\eta^2_p = .49$, suggérant que les participants étaient plus rapides pour détecter le phonème-cible lorsque l'item-cible était présenté en condition AV. Aucun effet d'interaction entre les 2 facteurs de l'analyse n'a pu être mis en évidence, $F_1 < 1$.

[36] Un indice d' a été également été calculé puis soumis à une ANOVA 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot). Un effet lexical a été obtenu, suggérant de meilleures performances pour détecter le phonème-cible dans un mot plutôt que dans un pseudo-mot, $F_1(1, 59) = 4.41$, $p < .05$, $\eta 2p = .07$. Aucun autre effet principal ni d'interaction entre les 2 facteurs de l'analyse ne s'est révélé significatif, tous les $F_1 < 1$.

3.3.3. Discussion

L'objectif principal de l'Etude 2 était de déterminer si, à l'instar de l'Etude 1, le fait de voir les mouvements articulatoires de son interlocuteur participe, en présence d'une information auditive congruente, au processus d'activation des unités lexicales. Pour tester cette hypothèse, nous avons utilisé une tâche de détection de phonèmes dans des mots et des pseudo-mots, en modalité A ou AV en l'absence (sans bruit) ou en présence de bruit blanc dans le signal acoustique (à -9 dB, et à -18 dB). A la différence de l'Etude 1, nous avons choisi des phonèmes vocaliques (plutôt que consonantiques) comme phonèmes-cibles, dans le but d'augmenter la saillance perceptive de ces derniers. Les voyelles sont généralement considérées, sur le plan perceptif, comme plus saillantes que les consonnes (Ladefoged, 2001) et comme résistant mieux à l'ajout de bruit dans le signal acoustique (Nooteboom & Doodeman, 1984, cité par Cutler et al., 2000). En effectuant cette modification, l'objectif était de réduire la difficulté générale de la tâche afin d'obtenir des mesures plus fines que le pourcentage de réponses correctes (i.e., mesures en temps réel, temps de réponse).

3.3.3.1.1. Apport de l'information visuelle

Les résultats de l'Etude 2 indiquent que, globalement, de meilleures performances ont été obtenues en modalité AV qu'en modalité A. En présence de bruit, ces données sont compatibles avec celles obtenues dans l'Etude 1 et dans la littérature (e.g., Sumby & Pollack, 1954, Benoît et al., 1994, cf. section 1.3.1.1). Les résultats mettent ensuite en évidence que, de manière identique à l'Etude 1, les participants étaient plus rapides pour détecter un phonème en modalité AV qu'en modalité A, lorsque le signal acoustique est détérioré. Dans l'Etude 1, nous avons observé que l'apport de l'information visuelle améliorait et accélérait le processus de détection des phonèmes-cibles surtout pour ceux les plus saillants sur le plan articulatoire (i.e., ayant une place d'articulation bilabiale). En l'absence de bruit, les phonèmes vocaliques étaient également plus rapidement détectés en condition AV qu'en modalité A. Ces données viennent donc renforcer l'idée qu'en présence de l'information auditive, le fait de percevoir les mouvements des articulateurs du visage de son interlocuteur facilite non seulement le processus de détection de phonèmes en situation bruitée mais *accélère* également ce dernier et ce même lorsque le signal acoustique est intact (voir Cox et al., 1999, pour des résultats similaires). Les résultats de l'Etude 2 montrent que la présence d'information visuelle permet, même lorsque l'information auditive est complètement intelligible (scores au plafond, $M = 98.1$ % en modalité auditive), de faciliter le processus de

détection de phonèmes, même lorsque le signal acoustique est intact. Ces données pourraient éventuellement concorder avec l'idée que le fait de voir les gestes articulatoires de notre interlocuteur nous permet d'anticiper leur conséquence acoustique (Cathiard, 1994 ; Jesse & Massaro, 2010 ; Munhall & Tohkura, 1998 ; Smeele, 1994). Cette idée sera plus clairement développée dans la section 6.2.2.1 du Chapitre 6.

3.3.3.1.2. Apport de l'information lexicale

Egalement en accord avec la majorité des données issues de la littérature (e.g., Frauenfelder et al., 1990) et avec celles issues de l'Etude 1, les résultats de l'Etude 2 mettent en évidence un effet de supériorité du mot, signifiant que les participants étaient plus rapides (excepté à en modalité A à -18 dB) et avaient des pourcentages de détections correctes plus élevés (excepté à en modalité A à -18 dB et en l'absence de bruit) lorsque le phonème vocalique cible était inséré dans un mot plutôt que dans un pseudo-mot. Bien que cet effet ne soit pas toujours répliqué dans la littérature avec ce type de tâche et varie par exemple en fonction de la place (dans le mot) du phonème à détecter (e.g., Frauenfelder et al., 1990, voir Connine & Titone, 1996 pour une revue), il semblerait que dans notre étude, le contexte lexical influence globalement le processus de détection des phonèmes.

3.3.3.1.3. Apport combiné de l'information visuelle et lexicale

Comme évoqué dans le paragraphe précédent, lorsque l'information auditive était intacte, les résultats de l'Etude 2 indiquent un effet de supériorité du mot sur les temps de réponse, mettant en évidence que les participants étaient plus rapides pour détecter un phonème dans un mot plutôt que dans un pseudo-mot. En l'absence de bruit, la taille de cet effet lexical ne différait pas significativement en fonction de la modalité de présentation de l'item-cible, comme nous l'avons également observé dans l'Etude 1. En d'autres termes, en l'absence de détérioration du signal acoustique, la taille de l'effet de supériorité du mot semble être semblable en modalité A (M_{effet} = 57 ms) et en modalité AV (M_{effet} = 56 ms). Ce résultat suggère que l'information visuelle participe au processus de reconnaissance de mots majoritairement lorsque celui-ci est difficile.

En présence de bruit (i.e., à -9 dB et à -18 dB), les résultats issus de l'Etude 2 diffèrent de ceux obtenus de l'Etude 1 sur deux points majeurs : (1) l'observation sur les pourcentages de réponses correctes d'un effet lexical en modalité A plus important qu'en modalité AV à -9 dB (2) l'obtention d'un effet lexical plus important en modalité AV qu'en modalité A sur les temps de réponse.

Concernant le point (1), les résultats indiquent, contrairement à nos hypothèses et aux données issues de l'Etude 1, un effet de supériorité du mot plus important en modalité A qu'en modalité AV à -9 dB. Cette donnée, apparemment incompatible avec nos hypothèses, peut être expliquée de deux manières. Premièrement, il est possible que plus le signal de parole est ambigu, plus le contexte lexical influence la perception des phonèmes. Supposer cela permettrait de prédire l'obtention d'un effet lexical plus important en modalité A qu'en modalité AV, l'intelligibilité du signal étant beaucoup plus faible dans la première condition que dans la seconde. De plus, l'effet lexical devrait augmenter à mesure que le signal acoustique est détérioré (i.e., plus le signal acoustique est ambigu). Or, à -18 dB, l'effet de supériorité du mot a uniquement été observé en modalité AV. Cette absence d'effet n'a pas pu être masquée par un effet « plancher » sur les pourcentages de réponses correctes puisque les performances, en modalité A à -18 dB, différaient significativement du hasard (situé à 50 % dans cette étude) pour les mots ($t(1, 59) = 2.55$, $p < .05$) comme pour les pseudo-mots ($t(1, 59) = 2.96$, $p < .005$). En conséquence, ce dernier résultat rend caduque cette première hypothèse interprétative.

La seconde supposition serait de dire que l'effet lexical pourrait partiellement être masqué par des performances « plafond » (M_{AV} = 92.7 % de réponses correctes) en modalité AV et non en modalité A à -9 dB (M_A = 73.4 %). Cela signifierait que l'information visuelle active les représentations lexicales mais que le bénéfice de celle-ci, dans le processus de l'accès au lexique, est essentiellement visible dès lors que la reconnaissance du mot est rendue assez difficile en modalité A (i.e., à -18 dB). Dans l'Etude 1, les résultats semblent également concordants avec cette hypothèse. En effet, l'effet lexical en condition AV était plus important et statistiquement plus robuste à -18 dB (M_{effet} = 10.8 %, $\eta^2_p = .24$, significatif par participants et par items) qu'à -9 dB (M_{effet} = 2.8 %, $\eta^2_p = .05$, seulement significatif par participants). Dans l'Etude 2, l'obtention d'un effet lexical à -18 dB seulement en modalité AV concorde également avec cette explication. En conclusion, ce résultat, de prime abord en contradiction avec nos hypothèses, pourrait tout de même supporter l'idée que l'information visuelle contribuerait principalement au processus d'activation des unités lexicales lorsque le processus de reconnaissance de mots est rendu difficile (voir Chapitre 4 et 6 pour plus de détails).

En conclusion, l'apport principal de l'Etude 2 est d'avoir mis en évidence un effet lexical sur les temps de réponse plus important en modalité AV qu'en modalité A en situation bruitée (point (2)). Ce résultat, permet d'apporter des données supplémentaires mesurées en temps réel, par rapport aux résultats issus de l'Etude 1 et de la littérature (e.g.,

Brancazio, 2004). En effet, cette dernière suggère que l'information visuelle contribue au processus de reconnaissance de mots per se en *accélérant* l'accès au lexique en présence d'une information auditive congruente et ce principalement lorsque l'information auditive est détériorée.

3.4. CONCLUSIONS

L'objectif de ce chapitre était d'étudier l'apport de l'information visuelle au processus de reconnaissance de mots, en présence d'une information auditive congruente et détériorée. A l'aide d'une tâche de détection de phonèmes nous avons étudié cette question en utilisant des phonèmes-cibles consonantiques (Etude 1) et vocaliques (Etude 2), insérés dans des mots ou des pseudo-mots, présentés en modalité AV ou A, avec différents niveaux de détérioration du signal acoustique (sans bruit, -9 dB, -18 dB). Premièrement, nous avons montré que le fait de voir le visage de son interlocuteur augmente non seulement l'intelligibilité des phonèmes en situation bruitée, mais accélère également leur détection (Etude 1 et 2). Ensuite, nous avons mis en évidence que la présence de ce signal visuel contribue à l'activation des représentations lexicales (Etude 1 et 2) en accélérant le processus d'accès au lexique lorsque le signal acoustique est détérioré (Etude 2). Ce résultat suggère que l'information visuelle contribuerait essentiellement au processus de reconnaissance de mot lorsque ce dernier est rendu difficile.

Néanmoins, certaines questions restent en suspens, concernant notamment le décours temporel des mécanismes impliqués dans ce processus. En effet, le design utilisé dans les Etude 1 et 2 ne nous permet pas de conclure si c'est le produit de l'intégration audiovisuelle (i.e., la fusion des informations auditives et articulatoires) ou le traitement du signal visuel, effectué séparément du signal acoustique, qui a permis d'accéder plus efficacement aux unités lexicales en modalité AV. Cette remarque soulève une question primordiale pour l'étude de la perception de la parole. En effet, depuis la mise en évidence de l'effet McGurk (McGurk & MacDonald, 1676), déterminer à quel moment, dans le décours temporel du processus de reconnaissance de mots, l'intégration auditive et visuelle s'effectue constitue un large débat dans la littérature (e.g., Schwartz et al., 1998). Certains supposent que cette fusion d'informations s'effectuerait très tôt dans le processus de reconnaissance de mots (e.g., Galantucci et al., 2006), alors que d'autres postulent que l'intégration audiovisuelle ne s'effectuerait que tardivement (e.g., Massaro & Chen, 2008). Nous discuterons de cette problématique dans la section 6.2.3 du Chapitre 6.

En lien avec cette question, une question peut également être formulée à la suite de ce travail. En effet, nous avons montré dans ce Chapitre que l'information visuelle, en présence d'une information auditive congruente, permet de contacter le lexique. Le principal objectif du Chapitre 4 consiste à évaluer le rôle spécifique de l'information visuelle *seule* dans le processus d'activation des unités lexicales.

CHAPITRE 4. ROLE DE L'INFORMATION VISUELLE SEULE DANS L'ACCES AU LEXIQUE

4.1. INTRODUCTION

Le but du Chapitre 4 est d'examiner l'apport de l'information visuelle seule dans le processus d'accès au lexique chez l'adulte. En d'autres termes, notre objectif est ici d'explorer si l'information visuelle *seule* permet d'activer les représentations lexicales, en l'absence de bruit ou de toute autre information auditive.

4.1.1. Travaux antérieurs

Plusieurs études ont cherché à évaluer la capacité d'individus à reconnaître des mots en modalité visuelle seule, notamment chez les malentendants (voir Bernstein, et al., 2000; Erber, 1974; Rouger, et al., 2007; Strelnikov, et al., 2009, parmi de nombreux travaux à ce propos). Dans cette section, nous allons examiner plus en détails celles qui se sont intéressées à évaluer les compétences des individus normo-entendants à identifier des mots isolés en modalité visuelle seule (i.e., capacités de labiolecture). Plus spécifiquement, nous allons, dans un premier temps, évoquer celles qui ont cherché à établir un lien entre cette compétence et les caractéristiques des stimuli inhérentes au niveau lexical, telles que la fréquence (e.g., Auer, 2002, 2009; Feld & Sommers, 2011; Kaiser, et al., 2003; Mattys, et al., 2002; Tye-Murray, et al., 2007, voir e.g., Auer, 2009, pour une revue).

Les travaux de Mattys et al. (2002), ont cherché à évaluer si la capacité à identifier des mots en modalité visuelle seule pouvait être modulée en fonction (1) de leur fréquence dans le langage oral et (2) de leur ressemblance perceptive (au niveau des gestes articulatoires) avec d'autres mots. Pour cela, les auteurs ont manipulé la fréquence de leurs stimuli mais également le nombre de mots partageant les mêmes visèmes consonantiques avec chaque stimulus (voir Chapitre 1 section 1.2.3, pour plus de détails à propos de la notion de visème). Cette quantité a été définie sous le terme de « Lexical Equivalence Class[37] », LEC. Il s'agit donc d'un paramètre d'ordre lexical, puisqu'il désigne le nombre de *mots* qui ressemblent visuellement au stimulus considéré. Dans cette étude, la moitié des stimuli disposait d'une LEC faible (N = 1) alors que l'autre moitié disposait d'une LEC importante (N > 10). Leurs résultats montrent que les participants normo-entendants avaient de meilleures performances pour identifier des mots isolés, présentés en modalité visuelle seule, lorsque ceux-ci étaient fréquents et avaient une LEC faible (voir e.g., Auer, 2009, Feld & Sommers, 2011, pour des résultats similaires). Les performances en lecture labiale étant sensibles à des paramètres inhérents au niveau lexical (fréquence des mots ou taille du LEC),

[37] Littéralement : « classe d'équivalence lexicale »

leurs résultats suggèrent donc que l'information visuelle seule permet d'activer les représentations lexicales.

Dans la même veine d'arguments, les travaux effectués par Kaiser et al. (2003) ont mis en évidence que les scores d'identification de mots en modalité visuelle seule étaient plus élevés lorsque ces mots étaient fréquents dans le langage et disposaient d'une densité de voisinage phonologique faible, autrement dit lorsque les mots étaient faciles à identifier (voir Tye-Murray et al., 2007, pour des résultats similaires). Globalement, ces résultats suggèrent que l'information visuelle permet d'activer les représentations lexicales, en l'absence d'information auditive. Cependant, tous ces travaux ont utilisé des tâches d'identification de mots en modalité visuelle seule. Or ce type de paradigme est sujet au développement d'un certain nombre de stratégies visant à *deviner* le mot prononcé (e.g., Lyxell & Rönnberg, 1987). Ainsi, la mise en place de stratégies conscientes pourrait également expliquer les résultats obtenus par l'ensemble de ces études, le pourcentage d'identifications correctes étant systématiquement plus élevé pour les mots les plus faciles à reconnaître et donc à deviner (i.e., fréquents et disposant d'une densité de voisinage phonologique faible). De fait, les études utilisant ce type de paradigme ne permettent pas de déterminer si les résultats observés sont (1) dus à l'activation *automatique* des représentations lexicales par l'information visuelle ou (2) imputables au développement de stratégies conscientes.

Indépendamment de ce débat, il apparaît également que ce type de tâche ne nous permet pas d'obtenir des mesures en temps réel quant à l'intervention de l'information visuelle dans le processus de reconnaissance de mots. Le paradigme d'amorçage est à l'inverse tout à fait adapté pour répondre à la question posée. En effet, ce type de design permet de mesurer des temps de réponse, évaluant en temps réel les processus impliqués dans une tâche. Ensuite, en faisant varier le type de relation (sémantique, phonologique, etc.) entre l'amorce et la cible, ce paradigme peut également mesurer l'impact de différents aspects linguistiques sur le processus étudié (e.g., Kim, Davis, & Krins, 2004). Enfin, les tâches d'amorçage permettent d'interroger des processus automatiques, ne pouvant pas forcément être mis en évidence par des tâches d'identification de mots du phénomène étudié (Forster & Davis, 1984).

A notre connaissance, trois études effectuées en anglais ont examiné le rôle de l'information visuelle seule dans le processus d'activation des représentations lexicales ; ces dernières ont toutes utilisé un paradigme d'amorçage par répétition (Buchwald, Winters, & Pisoni, 2009; Dodd, Oerlemens, & Robinson, 1989; Kim, et al., 2004). Ce type de paradigme consiste à présenter un stimulus entier en amorce puis ce même stimulus entier en tant que cible. La modalité de présentation de l'amorce est souvent différente de celle de la cible (e.g.,

amorce auditive, cible écrite). Cette manipulation a pour but d'étudier si deux productions d'un même stimulus, issues d'une modalité différente permettent tout aussi bien d'activer la représentation en mémoire correspondant à ce stimulus.

L'objectif de Dodd et al. (1989) était d'étudier si la présentation d'un ensemble de mots en modalité visuelle seule (gestes articulatoires du visage du locuteur en l'absence de son) facilitait la catégorisation sémantique ultérieure de ces mêmes mots, présentés en modalité auditive. Pour cela, dans une première phase (« phase d'amorçage »), un ensemble de 10 mots était présenté aux participants, en modalité visuelle seule. Lors de cette étape, les participants avaient pour tâche de déterminer si le mot prononcé correspondait à un nom d'animal (N = 5) ou de plante (N = 5). Dans la seconde partie de l'expérience, les mêmes 10 mots « familiers » étaient présentés en modalité auditive. Dix autres mots (5 d'animaux et 5 de plantes) « nouveaux » (i.e., étant connu par les participants mais n'ayant pas été montrés en phase 1) étaient également présentés. La tâche demandée aux participants était la même que dans la phase précédente. Les résultats mettent en évidence des temps de réponse significativement plus courts dans la seconde phase pour catégoriser les items « familiers » que pour les mots « nouveaux ». Cette étude suggère donc que les stimuli « familiers » présentés en phase d'amorçage en modalité visuelle seule ont activé le même réseau lexico-sémantique que ces mêmes items présentés en modalité auditive dans la seconde phase, facilitant leur catégorisation dans cette étape. Ainsi ce résultat pourrait éventuellement suggérer que l'information visuelle seule permet d'activer les représentations lexicales. Cependant, la critique majeure que l'on peut effectuer à l'égard de cette étude concerne le paradigme utilisé. En effet, les participants devaient effectuer lors de la première phase une tâche sur les mots familiers en modalité visuelle seule ce qui les a probablement incités à deviner le mot en question. Ainsi, ce paradigme, même s'il est désigné sous le terme d' « amorçage par répétition » ne permet pas de déterminer si les résultats observés sont dus à l'activation automatique des représentations lexicales ou uniquement imputables au développement de stratégies.

Kim et al. (2004) ont proposé une étude utilisant également le paradigme d'amorçage par répétition, mais ne présentant pas ce type de problème méthodologique. Celle-ci consistait à présenter une amorce en modalité visuelle seule (gestes articulatoires du visage du locuteur en l'absence de son) suivie d'une cible présentée à l'écrit ou en modalité auditive seule. L'amorce était soit un mot (e.g., /bak/, «back », dos) soit un pseudo-mot (e.g., /nʌnθ/, «nunth »). Elle pouvait également être identique à la cible (condition reliée, e.g., présentation écrite ou auditive de « back » pour la condition mot ; « nunth » pour la condition pseudo-mot) soit différente de la cible (condition non reliée, e.g., /ʃarp/, « sharp »,

coupant, pour le mot-cible « back » ; « scay », /skeɪ/, pour le pseudo-mot-cible « nunth »). Les auteurs ont mis en évidence que les participants étaient plus rapides pour dire que la cible était un mot ou un pseudo-mot (tâche de décision lexicale) ou pour la dénommer (tâche de dénomination) lorsque celle-ci était précédée d'une amorce reliée plutôt que non reliée. Or, cet effet d'amorçage facilitateur était significatif uniquement lorsque les stimuli étaient des mots. Comme aucun effet n'a pu être mis en évidence pour les pseudo-mots, ces résultats indiquent que l'articulation silencieuse pour un mot *entier* (e.g., « back ») a permis d'activer sa représentation lexicale, accélérant sa reconnaissance ultérieure. De plus, comme cette facilitation a été retrouvée que la cible soit présentée à l'écrit (modalité visuelle) ou en modalité auditive seule, ces résultats suggèrent que les processus impliqués dans cet effet sont de nature amodale.

Quelques années plus tard, Buchwald et al. (2009) ont pu mettre en évidence qu'une cible présentée en modalité auditive avec du bruit (e.g., /beɪk/, « bake », faire cuire) était plus facilement identifiée (tâche d'identification de mots) lorsque celle-ci était précédée par l'articulation silencieuse du même mot (condition reliée, e.g., /beɪk/, « bake ») plutôt que par la présentation d'une image fixe du visage de la locutrice (condition non reliée). Afin d'analyser plus en détails leurs résultats, les auteurs ont également mesuré le nombre de phonèmes de la cible correctement identifiés pour la condition reliée par rapport à la condition non reliée. Leurs résultats mettent en évidence un effet d'amorçage sur cette variable, indiquant que le nombre de segments correctement identifiés était plus important en condition reliée que non reliée. Egalement, cet effet d'amorçage était plus important pour des mots-cibles de haute fréquence lexicale et disposant d'une densité de voisinage phonologique faible que pour des mots de basse fréquence lexicale, disposant d'une densité de voisinage phonologique importante. Ces résultats suggèrent donc que l'apport de l'information visuelle serait plus important pour reconnaître un mot en modalité auditive lorsque celui-ci est facile à identifier, c'est-à-dire lorsqu'il est fréquent dans le langage oral et ne dispose de similarités acoustico-phonologiques qu'avec peu d'autres mots. Ces paramètres (i.e., densité de voisinage phonologique et fréquence d'occurrence) étant connus pour influencer le processus de reconnaissance de mot à un niveau lexical, ces données suggèrent que le traitement de l'information visuelle seule engendre une facilitation à ce niveau et ce principalement pour reconnaître des mots fréquents ayant peu de compétiteurs (i.e., densité de voisinage phonologique faible).

4.1.2. Objectifs et méthodes des Etudes 3 et 4

L'objectif de ce chapitre consiste à examiner de manière plus approfondie le rôle joué par l'information visuelle seule dans l'accès au lexique. Si le fait de voir les mouvements oro-faciaux de notre interlocuteur d'un mot permet d'activer sa représentation lexicale (e.g., Kim et al., 2004), nous faisons l'hypothèse que cette facilitation interviendrait *précocement* dans le processus de traitement de la parole. En d'autres termes, nous postulons que l'information visuelle jouerait un rôle prépondérant dans le processus initial d'activation des candidats lexicaux. Ainsi, le geste articulatoire correspondant à la première syllabe d'un mot (et non pas à l'articulation silencieuse d'un mot entier) devrait fournir suffisamment d'information pour faciliter sa reconnaissance. Afin de tester cette hypothèse, nous avons décidé d'utiliser un paradigme d'amorçage phonologique partiel (voir Chapitre 2, section 2.2.4, pour plus de détails à propos de ce paradigme). Initialement, le paradigme d'amorçage phonologique a été utilisé afin d'examiner l'organisation des représentations phonologiques dans le lexique (e.g., Hamburger & Slowiaczek, 1996; Marslen-Wilson & Zwitserlood, 1989; Radeau, et al., 1989; Slowiaczek & Hamburger, 1992; Slowiaczek, et al., 1987; Spinelli, 1999; Spinelli, et al., 2001; Zwitserlood, 1989). Ce type de design expérimental permet d'étudier la dynamique de la reconnaissance des mots parlés. Le paradigme d'amorçage phonologique *partiel* a notamment permis de mettre en évidence l'importance du début de mot dans le processus d'accès au lexique en modalité auditive seule (e.g., Marslen-Wilson & Zwitserlood, 1989; Spinelli, 1999; Spinelli, et al., 2001; Zwitserlood, 1989). Spinelli et collègues (2001) ont par exemple montré que la présentation d'une amorce auditive correspondant à la première syllabe d'un mot-cible écrit (e.g., /vɛʁ/, « ver » → /vɛʁvɛn/, « VERVEINE ») facilitait sa reconnaissance, en comparaison avec la présentation d'une amorce non reliée (e.g., /kw$\tilde{\varepsilon}$/, « coin »). Cependant, aucun effet facilitateur n'a pu être observé lorsque l'amorce auditive partageait la dernière syllabe avec la cible (e.g., /vɛn/, « veine » → /vɛʁvɛn/, « VERVEINE »). Ces résultats mettent en évidence l'importance du début de mot dans le processus d'activation des candidats lexicaux (e.g., le modèle de la Cohorte II, e.g., Marlsen-Wilson, 1987, 1990). Plus spécifiquement, ces données montrent que la première syllabe d'un mot (e.g., « ver ») comporte suffisamment d'information auditive pour activer la représentation du mot « VERVEINE » et ainsi faciliter (i.e., accélérer) sa reconnaissance ultérieure.

L'objectif de la prochaine expérimentation (Etude 3) consiste à explorer si une telle facilitation est également observée avec une amorce présentée en modalité visuelle seule. En d'autres termes, cette étude vise à examiner si le *geste articulatoire* correspondant à la

première syllabe d'un mot (amorce) est suffisant pour activer sa représentation lexicale et ainsi accélérer sa reconnaissance ultérieure (cible). Si tel est le cas, le second but de cette expérimentation consiste à comparer la facilitation observée en présence d'une information visuelle seule à celle obtenue en modalité auditive et audiovisuelle. Pour cela, nous avons utilisé un paradigme d'amorçage phonologique partiel similaire à celui de Spinelli et al. (2001), excepté que nous avons fait varier la modalité de présentation de l'amorce selon trois conditions : auditive seule (A), visuelle seule (V) et audiovisuelle (AV). La cible, elle, est toujours présentée en modalité auditive seule. Si le fait de voir le premier geste articulatoire d'un mot permet effectivement d'en activer sa représentation, nous devrions obtenir un effet d'amorçage en modalité visuelle seule. Ensuite, si ce bénéfice est également significatif en présence d'une information auditive non détériorée, nous devrions obtenir une facilitation plus importante en condition AV que A. Enfin, conformément aux études précédentes (e.g., Spinelli et al., 2001), nous devrions obtenir un effet d'amorçage en condition V.

4.2. ETUDE 3 : VOIR LE GESTE ARTICULATOIRE CORRESPONDANT A LA PREMIERE SYLLABE D'UN MOT FACILITE-T-IL SA RECONNAISSANCE ?

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L. & Spinelli, E. (en révision). Visual speech facilitates the early phases of word recognition: Evidence from fragment priming tasks. *Language and Cognitive Processes.*

4.2.1. Méthode

4.2.1.1. Participants

Soixante-trois participants (dont 42 femmes et 11 hommes) âgés de 17 à 38 ans (M = 22 ans) ont été recrutés pour cette étude. Tous étaient de langue maternelle française. Aucun d'entre eux n'avait de déficit sur le plan auditif et tous avaient une vision normale ou corrigée. La majorité des participants était étudiants en Psychologie à l'Université Pierre Mendès France de Grenoble et recevaient un bon d'expérimentation en échange de leur participation.

4.2.1.2. Stimuli

4.2.1.2.1. Mots

Un corpus de 90 mots bisyllabiques (e.g., /by.ʁo/, « bureau », le point indiquant la frontière syllabique, a été sélectionné en utilisant la base de données LEXIQUE 2 (New et al., 2001, cf. Annexe C). Chacun de ces items commençait par une syllabe de type CV qui était toujours composée par une consonne articulée à l'avant du conduit vocal (/b/, /m/ /p/, /f/, /v/, /s/) et par une voyelle protruse (i.e., provoquant un arrondissement des lèvres : /u/, /o/, /y/). Cette précaution a été prise afin de maximiser la visibilité des amorces sur le plan articulatoire. Chacun d'entre eux était associé avec deux amorces monosyllabiques de type CV, entretenant (condition reliée : e.g., /by/-/byʁo/) ou n'entretenant pas (condition non reliée : e.g., /fo/-/byʁo/) une relation de recouvrement initial (partage de la première syllabe) avec le mot-cible. La syllabe pouvant constituer une unité fonctionnelle pour l'accès au lexique (e.g., Mehler, et al., 1981; Spinelli, 1999) nous nous sommes assurés que l'alignement syllabique entre l'amorce et la première syllabe de la cible était identique pour chaque triplet amorce reliée/non reliée/cible (e.g., /by/-/by.ʁo/ et non /byʁ/-/ by.ʁo /). La durée moyenne des mots-cibles était de 508 ms. La fréquence moyenne des mots était de 24.64 opm.

4.2.1.2.2. Pseudo-mots

La tâche demandée aux participants étant une tâche de décision lexicale (i.e., décider si le stimulus entendu est un mot ou un pseudo-mot). Quatre-vingt-dix pseudo-mots bisyllabiques (e.g., /by.ʁat/) ont été créés à partir des 90 mots décrits ci-dessus, en respectant les mêmes caractéristiques de sélection. Ainsi, chacun d'entre eux était également associé à une amorce reliée (e.g., /by/-/ byʁat/) et non reliée (e.g., /fo/-/byʁat/). Chaque item entretenait également un lien de recouvrement initial avec l'amorce, afin que les participants n'établissent pas leur jugement de lexicalité sur la seule base d'une liaison entre amorces et cibles.

4.2.1.2.3. Items de remplissage

Cent quatre-vingt items bisyllabiques de remplissage (90 mots et 90 pseudo-mots) non reliés avec l'amorce ont également été sélectionnés (e.g., /kanaʁ/, « canard », /ʃãfi/, etc.). Par conséquent, la proportion d'items reliés étaient de 25 % seulement, afin de limiter

le développement de stratégies de réponse par les participants (e.g., Hamburger & Slowiaczek, 1996).

4.2.1.2.4. Enregistrement des stimuli

L'ensemble des stimuli (amorces et cibles) a été enregistré dans une chambre sourde avec le même matériel et dans les mêmes conditions que l'Etude 1. La locutrice était la même que celle de l'Etude 2. Cette dernière était placée devant un fond vert et seulement sa tête, son cou et le haut de ses épaules étaient visibles (cf. Figure 26). Pour les besoins spécifiques de cette étude, chaque stimulus a également été segmenté de manière à ce que le délai entre la fin de l'amorce et le début de la cible (Intervalle Inter- Stimuli, ISI) soit de 50 ms et que le SOA (« Stimulus Onset Asynchrony », i.e., le délai entre le début de l'amorce et le début de la cible) n'excède pas 250 ms (Hamburger & Slowiaczek, 1996).

4.2.1.3. Procédure

Les participants étaient placés dans une pièce calme, à 50 cm d'un écran DELL de 19 pouces (1024 x 768 pixels) et entendaient les stimuli par l'intermédiaire d'un casque SENHEISER HD 212 pro. La composante vidéo des stimuli était présentée à une fréquence de 25 images/seconde. Le signal acoustique était présenté à une fréquence d'échantillonnage de 44100 Hz. Chaque essai commençait par un point de fixation. L'amorce était ensuite présentée, suivie 50 ms plus tard par la présentation d'une cible auditive. Les participants avaient pour tâche de décider le plus rapidement possible si la cible était un mot ou non en appuyant sur une des deux touches situées de part et d'autre du clavier (i.e., tâche de décision lexicale). La main dominante était désignée pour la réponse « mot ». L'expérience se déroulait selon trois blocs de 120 items chacun, chacun correspondant à une modalité de présentation de l'amorce : A, V ou AV. L'amorce (en modalité A) ainsi que la cible était toujours accompagnée de l'image fixe du visage de la locutrice. L'ordre de présentation de chacun de ces blocs était aléatoire. A l'intérieur de chaque bloc, le statut lexical des cibles (mot vs. pseudo-mot) ainsi que le type d'amorce (reliée vs. non reliée) étaient également présentés dans un ordre aléatoire. Tout au long de l'expérience, les participants avaient pour consigne de prêter attention tout aussi bien à la composante acoustique que vidéo des stimuli. Pour chaque item cible, la condition de présentation de l'amorce ainsi que la nature du lien entre l'amorce et la cible étaient contrebalancées entre les participants. Chaque paire de mot/pseudo-mot était distribuée selon 6 listes correspondant aux 6 conditions de présentation des amorces : AV reliée ; AV non reliée ; A relié ; A non reliée ; V Reliée ; V non reliée. Chaque item (expérimental ou de remplissage) était donc présenté une seule fois à

chaque participant. Une période d'entraînement de 10 items précédait chaque nouvelle phase. Le déroulement de chaque essai en fonction des conditions est représenté dans la Figure 26. La génération des stimuli et la collecte du type et du temps de réponse était assurée par le logiciel E-Prime 2.0 (*Psychological Software Tools, Pittsburgh, PA*). La totalité de la passation durait 25 minutes environ.

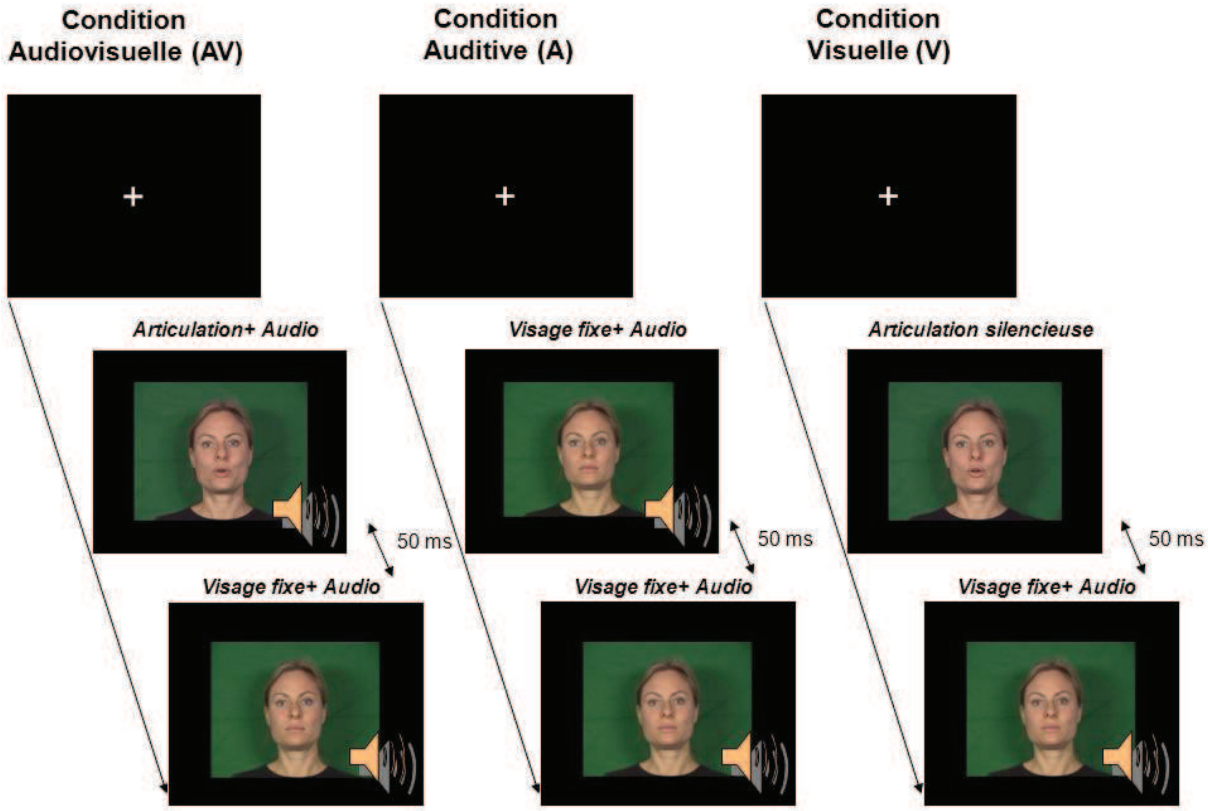


Figure 26. *Représentation schématique des différentes conditions expérimentales de l'Etude 3. Pour chaque condition de présentation de l'amorce, celle-ci pouvait soit être reliée (/by/-/byʁo/) ou non reliée avec la cible (/fo/-byʁo).*

4.2.2. Résultats

Le pourcentage d'erreurs ainsi que la moyenne des temps de réponse (mesurés à partir du début du mot-cible, uniquement sur les réponses correctes pour les mots[38]) ont été calculés pour chaque participant et chaque paire d'items. Deux items ont été retirés de l'analyse, ces derniers ayant un pourcentage d'erreur supérieur à 30 %. Une analyse de la variance (ANOVA) 3 (Modalité de Présentation : AV vs. A vs. V) x 2 (Type d'Amorce :

[38] Une analyse de la variance (ANOVA) 3 (Modalité de Présentation : AV vs. A vs. V) x 2 (Type d'Amorce : Reliée vs. Non Reliée) a été effectuée par participants (F1) et par items (F2) sur les temps de réponses effectués sur les réponses correctes mais également sur le pourcentage d'erreurs observés sur les pseudo-mots. Aucun effet principal ni d'interaction ne s'est révélé significatif.

Reliée vs. Non Reliée) a donc été effectuée en échantillon appariés sur 63 participants, par participants (F_1) et par items (F_2), sur les temps de réponse. Les temps de réponse considérés comme aberrants (i.e., inférieur à 100 ms ou supérieur à 2500 ms) ou se situant à plus de 2 écart-types de la moyenne du participant pour chaque condition respective ont été exclus de l'analyse. Cette opération a écarté 2.3 % des données totales. Le pourcentage d'erreurs étant relativement faible (< 1 %) aucune analyse statistique n'a été conduite sur cette mesure. L'ensemble des résultats pour les conditions de l'Etude 3 est présenté dans le Tableau 10.

Tableau 10. *Temps de réponse (en ms) et pourcentage d'erreurs (en gris) en fonction des différentes conditions de l'Etude 3. L'erreur type est présentée entre parenthèses.*

Modalité de l'amorce	Type d'amorce		Effet d'amorçage
	Reliée	Non Reliée	
AV	812 (9.2)	855 (10.1)	43
A	827 (10)	860 (10.6)	34
V	832 (8.6)	846 (9.6)	14

L'analyse sur les temps de réponse a tout d'abord révélé un effet principal du type d'amorce, $F_1(1, 62) = 64.96$, $p < .001$, $\eta^2_p = .51$, $F_2(1, 87) = 71.08$, $p < .001$, $\eta^2_p = .45$, mettant en évidence que les participants étaient plus rapides pour reconnaître un mot dans la condition reliée que non reliée (effet d'amorçage facilitateur). Un effet d'interaction entre la modalité et le type d'amorce a été obtenu, $F_1(2, 124) = 4.3$, $p < .05$, $\eta^2_p = .07$, $F_2(2, 174) = 4.23$, $p < .05$, $\eta^2_p = .05$. Des comparaisons planifiées ont révélé que l'effet d'amorçage était significatif lorsque l'amorce était présentée en modalité visuelle, $F_1(1, 62) = 3.93$, $p = .05$, $\eta^2_p = .06$, $F_2(1, 87) = 4.27$, $p < .05$, $\eta^2_p = .05$, mais que la taille de cette facilitation était plus importante pour la condition auditive et audiovisuelle, $F_1(1, 62) = 8.57$, $p = .005$, $\eta^2_p = .12$, $F_2(1, 87) = 7$, $p < .01$, $\eta^2_p = .075$. Aucune différence d'amorçage n'a été mise en évidence entre la modalité auditive et audiovisuelle, $F_1(1, 62) < 1$.

4.2.3. Discussion

L'objectif de cette étude était d'examiner si le fait de voir le geste articulatoire correspondant à la première syllabe d'un mot (amorce) constitue une information suffisante pour activer sa représentation lexicale et ainsi faciliter sa reconnaissance ultérieure (cible).

Premièrement, les résultats indiquent que les mots cibles étaient plus rapidement reconnus lorsque ceux-ci étaient précédés par une amorce reliée que par une amorce non

reliée. Cet effet d'amorçage facilitateur a été observé lorsque l'amorce était présentée en condition A, AV mais également lorsque seul le geste articulatoire correspondant à la première syllabe de la cible était proposé (condition V). Aucun effet significatif n'a été observé lorsque les cibles étaient des pseudo-mots. Cela *suggère* que le locus de l'effet observé était lexical plutôt que pré-lexical, les pseudo-mots ne pouvant être décodés au niveau lexical. Cette facilitation obtenue pour les mots est généralement interprétée comme le fait que l'information contenue dans l'amorce (e.g., /by/) est suffisante pour activer les représentations lexicales partageant le même début (e.g., « bureau », « burin », etc.). L'ISI entre la fin de l'amorce et le début de la cible étant très court (50 ms), l'activation des unités lexicales n'aurait pas le temps de se dissiper entièrement avant l'arrivée de la cible (Spinelli, et al., 2001). C'est cette activation résiduelle qui permettrait de faciliter la reconnaissance du mot-cible lorsque celui-ci a été précédé par une amorce partageant la même syllabe initiale ou les mêmes phonèmes initiaux (condition reliée ; e.g., /by/ - /byʁo/, « bureau »), par rapport à une condition non reliée (/fo/ - /byʁo/, « bureau »). Ces résultats nous permettent de répliquer les résultats obtenus précédemment par Spinelli et al. (2001) pour la condition A, mais également de montrer que la présentation du visage de la locutrice articulant l'amorce facilite le traitement de la cible auditive. En d'autres termes, cette étude montre que le fait de voir l'articulation silencieuse pour /by/ accélère la reconnaissance du mot « bureau ». L'effet d'amorçage facilitateur obtenu pour la condition visuelle seule *suggère* que le fait de voir le visage de son interlocuteur produire le début de ce mot fournit suffisamment d'information pour permettre d'activer sa représentation lexicale. En conséquence, cela pourrait signifier que les traits articulatoires visibles (i.e., telles que la place d'articulation) contenues dans l'amorce en condition V constituent suffisamment d'information pour atteindre le niveau lexical. Cela *suggère* que conformément à nos hypothèses, l'information visuelle interviendrait dans les étapes précoces du processus de reconnaissance de mots, c'est-à-dire qu'elle participerait à l'étape d'activation des différents candidats lexicaux.

Ensuite, les résultats indiquent également que lorsque l'amorce était présentée en condition V, moins de facilitation a été observée par rapport aux conditions AV et A. Deux explications semblent pouvoir expliquer ce phénomène. Une première raison liée à la nature de l'information visuelle permettrait en effet de rendre compte de ce phénomène, mais également le fait que l'effet d'amorçage observé en présentation AV de l'amorce n'est pas significativement plus important que celle observée en modalité A. Le pattern de facilitation $V < A \approx AV$ peut venir du fait que le geste articulatoire de l'amorce utilisée en condition visuelle seule pourrait correspondre à plus de deux phonèmes (notion de visème, Fisher, 1968). Par exemple, les phonèmes /b/ et /p/ se distinguent sur le plan acoustique

uniquement par la présence de voisement, lié à une vibration plus ou moins importante des cordes vocales. Visuellement, cette caractéristique est très difficilement perceptible (e.g., Summerfield, 1987, 1991). Il en est de même pour la nasalisation, qui permet de distinguer un /b/ d'un /m/. En effet, ce trait résulte de la mise en action du voile du palais, qui n'est pas visible en situation de communication face à face. De la même manière, la différence entre les phonèmes /y/ et /u/ réside dans le fait que la position de la langue est plus recourbée vers l'arrière du conduit vocal pour un /u/. Ces deux voyelles entraînant une forte protrusion des lèvres, le mouvement de la langue est extrêmement difficile à percevoir visuellement. Ainsi, le mouvement articulatoire pour la syllabe /by/ est donc dans le signal visuel de parole extrêmement proche de celui pour /py/ et /my/, mais aussi de celui pour /bu/, /pu/ et /mu/. En accord avec cette hypothèse, la présentation de l'amorce /by/ permettrait tout aussi bien d'activer l'ensemble des représentations de mots commençant par /by/ comme « bureau » mais également ceux commençant par /py/, /my/, /bu/, /pu/ et /mu/ tels que « purée », « boulet » , « mulet », « poulet » et « moutarde ». La base de données LEXIQUE 3.71 (New, Pallier, Brysbaert, & Ferrand, 2004) indique que 250 mots français commencent par la syllabe /by/ dans la langue française. Ce nombre pourrait correspondre au nombre de candidats activés par la présentation auditive de /by/. Si le geste articulatoire pour /by/ permet bien d'activer tous les mots commençant par /by/ mais également par /py/, /my/ /bu/, /pu/ et /mu/, la taille de la cohorte ne serait plus de 250 mais de 250 (/by/) + 904 (/bu/) + 384 (/py/) + 525 (/my/) + 427 (/pu/) + 360 (/mu/) = 2850 candidats. Ainsi, selon cette hypothèse, il serait possible qu'un plus grand nombre de candidats lexicaux soit activé pour la condition V par rapport aux conditions A et AV, augmentant de ce fait la compétition lexicale et diminuant ainsi la taille de l'effet d'amorçage facilitateur dans le premier cas. Pour les mêmes raisons, cette explication peut également rendre compte du fait qu'en présence d'une information auditive clairement audible (i.e., dans la condition AV), le bénéfice de l'information visuelle soit de trop petite taille pour qu'une différence d'amorçage significative soit observée par rapport à la condition A. Cette idée permettrait donc d'expliquer également que bien que la taille de la facilitation observée pour la condition AV (M = 43 ms) soit descriptivement plus grande que celle observée en condition A (M = 34 ms) mais que cette différence n'est pas significative sur le plan statistique.

Deuxièmement, puisque l'effet d'amorçage obtenu en modalité A et AV sur une cible auditive est plus important que celui observé en modalité V, cela suggère que l'effet d'amorçage est dépendant de la concordance de la modalité de l'amorce par rapport à celle de la cible. Selon la logique des études intermodales (e.g., Kim et al., 2004), si l'effet d'amorçage

est modulé par ce facteur, cela suggère qu'un certain coût est observé pour le traitement selon un code commun, de l'amorce avec la cible. Pour les modèles de perception de la parole, ce raisonnement vient notamment soulever des questions concernant la nature des unités impliquées (amodale vs. spécifique à la modalité) permettant de questionner indirectement à quel moment, dans le décours temporel du processus d'accès au lexique, l'intégration des informations auditives et visuelles s'effectue. Cette question est soulevée si l'on suppose que la facilitation observée en condition V dans cette étude est plus faible qu'en modalité auditive car elle liée à un coût de traitement, lui-même suggérant l'existence d'une étape supplémentaire permettant de comparer les informations contenues dans l'amorce et dans la cible dans un code commun. Ce coût de traitement au niveau lexical signifierait que plusieurs unités ou plusieurs codages (auditif, visuel, etc.) existent pour un même mot. Le cas échéant, cela impliquerait que l'intégration audiovisuelle s'effectuerait après avoir contacté les unités lexicales (voir section 6.2.3.3 du Chapitre 6, pour plus de discussion à ce sujet).

4.2.4. Conclusions

La principale contribution de l'Etude 3 à l'étude de notre question est d'avoir mis en évidence que la présentation d'un geste articulatoire correspondant à la première syllabe d'un mot bisyllabique (amorce, e.g., /by/ pour /by.ʁo/) accélère sa reconnaissance ultérieure (cible). Cette facilitation est généralement interprétée comme le fait que l'information contenue dans l'amorce (e.g., /by/) est suffisante pour activer les représentations lexicales partageant le même début (e.g., « bureau », « burin », etc.). Cette activation n'ayant pas le temps de se dissiper entièrement avant l'arrivée de la cible, l'activation résiduelle permettrait de faciliter sa reconnaissance ultérieure lorsque celle-ci entretient une relation de recouvrement initial avec l'amorce (Spinelli, et al., 2001) pour une interprétation similaire avec de l'amorçage inhibiteur). L'originalité de ce travail est qu'il *suggère* que l'information visuelle contenue dans le geste articulatoire pour le début d'un mot est suffisante pour contacter les représentations lexicales. Toutefois, le design expérimental utilisé dans cette étude ne nous permet pas de conclure de manière *formelle* quant au locus (pré-lexical, lexical ou post-lexical) des mécanismes responsables de la facilitation observée dans cette étude. En d'autres termes, sur la base de ces résultats, il ne nous est pas possible d'infirmer l'hypothèse que l'effet d'amorçage notamment observé en modalité visuelle seule résulte de l'intervention de processus uniquement pré-lexicaux. Bien que nous nous soyons placés dans des conditions limitant le développement de stratégies de réponses (processus post-lexicaux) avec le paradigme d'amorçage phonologique (i.e., 25 % d'items reliés et 50 ms d'ISI,

Hamburger & Slowiaczek, 1996), la tâche utilisée dans cette étude ne nous permet pas d'écarter l'hypothèse que cette facilitation soit uniquement due à l'activation des unités situées avant le niveau lexical. Contrairement au paradigme de détection de phonèmes employé dans les Etudes 1 et 2 (cf. Chapitre 3), la tâche de décision lexicale ne permet pas de comparer les réponses obtenues pour les mots (impliquant une réponse positive « oui, c'est un mot ») avec les réponses pour les pseudo-mots (impliquant une réponse négative « non, ce n'est pas un mot »). Or, la comparaison mot versus pseudo-mot est couramment employée dans la littérature afin de déterminer si un effet est dû à l'intervention de mécanismes pré-lexicaux ou lexicaux, les pseudo-mots ne pouvant être décodés au niveau lexical. En effet, il est communément admis que des processus supplémentaires d'inhibition pourraient être à l'œuvre pour effectuer une réponse négative (i.e., sur un pseudo-mot) par rapport à la condition mot, où une réponse positive est demandée. Dans cette hypothèse, il est alors possible que dans notre étude, des effets d'amorçage soient présents sur les pseudo-mots mais que ces différences soient masquées par les processus d'inhibition également mis en jeu. En conséquence, il est possible que les effets d'amorçage observés dans l'Etude 3 soient liés à l'intervention de mécanismes pré et/ou post-lexicaux plutôt que lexicaux.

4.2.5. Objectifs de l'Etude 4

Si l'information visuelle seule permet d'activer les unités lexicales, la facilitation obtenue pour la condition V de l'Etude 3 devrait être modulée en fonction de paramètres lexicaux tels que la fréquence lexicale du mot-cible (e.g., en modalité auditive : Davis, Kim, & Barbaro, 2010; Dufour & Peereman, 2003, 2004). En effet, dans les modèles tels que TRACE (McClelland & Elman, 1986) ou Cohorte II (Marslen-Wilson, 1987, 1990) la fréquence des mots est considérée comme un paramètre influençant directement leur reconnaissance au niveau lexical. Par exemple, pour les modèles TRACE et celui de la Cohorte II, ce facteur définit un niveau d'activation de base plus important pour des unités lexicales correspondant à des mots fréquents que non fréquents. Cela permet d'expliquer qu'un mot fréquent est plus rapidement reconnu qu'un mot rare par des mécanismes situés directement au niveau lexical. Ainsi, plus une unité code pour un mot fréquent dans le langage, moins celle-ci nécessite de recevoir d'activation pour être reconnue. Cependant, certains auteurs supposent que les effets de fréquence ne seraient pas directement issus du niveau lexical (e.g., NAM, Luce & Pisoni, 1998). Ainsi, le modèle NAM postule que la fréquence du mot serait un paramètre situé à un niveau décisionnel post-lexical et que celui-ci viendrait biaiser le niveau d'activation des unités lexicales après que celles-ci aient été activées par l'input sensoriel.

Néanmoins, ces derniers postulent que ce paramètre interviendrait au niveau lexical avant que l'accès au lexique ne soit complété, permettant à l'effet de fréquence de posséder un locus lexical, plutôt que post ou pré-lexical. Ainsi, si l'effet d'amorçage observé pour l'Etude 3 est dû à l'intervention de mécanismes lexicaux, celui-ci devrait être modulé en fonction de la fréquence de la cible. L'objectif de l'Etude 4 consiste donc à examiner l'impact de la fréquence lexicale du mot-cible sur l'effet de facilitation obtenu par la présentation d'une amorce en modalité visuelle seule.

4.3. ETUDE 4 : APPORT DE L'INFORMATION VISUELLE SEULE DANS LE PROCESSUS DE RECONNAISSANCE DE MOTS : UNE FACILITATION LEXICALE ?

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L. & Spinelli, E. (en révision). Visual speech facilitates the early phases of word recognition: Evidence from fragment priming tasks. *Language and Cognitive Processes.*

4.3.1. Méthode

4.3.1.1. Participants

Vingt participants (dont 15 femmes et 5 hommes) âgés de 20 à 31 ans ($M = 25$ ans) ont été recrutés pour cette étude. Tous étaient de langue maternelle française. Aucun d'entre eux n'avait de déficit sur le plan auditif et tous avaient une vision normale ou corrigée. Aucun d'entre eux n'avait participé à l'Etude 3.

4.3.1.2. Stimuli

4.3.1.2.1. Mots

Un corpus de 30 paires de mots bisyllabiques de haute versus basse fréquence lexicale (e.g., /po.ze/, « poser » vs. /po.te/, « potée », le point indiquant la frontière syllabique, cf. Annexe D) ont été sélectionnés en utilisant la base de données « Freqfilm2 » de LEXIQUE 3.71 (New, et al., 2004). Chacun de ces items commençait par une syllabe de type CV. La moitié des items étaient considérés comme des mots de haute fréquence lexicale (F = 124.61 opm, intervalle, 26.8-626 opm), alors que la seconde partie correspondait à des mots de basse fréquence lexicale dans le langage oral (F = 0.78 opm, intervalle, 0-3.65 opm). La différence de fréquence lexicale entre ces deux groupes était statistiquement significative, $t(58) = 27.29$, $p < .001$. Nous avons vérifié que la durée de production entre les mots de haute fréquence (M = 482 ms) et de basse fréquence (M = 477 ms) n'était pas significativement différente, $t(58) < 1$. Egalement, nous nous sommes assurés que tous les items comportaient un point d'unicité phonologique au quatrième phonème, (excepté pour « vécu », dont le point d'unicité se situait au troisième phonème), ces facteurs pouvant augmenter « artificiellement » les effets de fréquence (e.g., Goldinger, 1996; Marslen-Wilson, 1990). Chaque paire de mot (haute fréquence-basse fréquence) était associée avec deux amorces monosyllabiques de type CV, entretenant (condition reliée : e.g., /po/-/poze/ et /po/-/pote/) ou n'entretenant pas (condition non reliée : e.g., /ʃi/-/ poze/ et /ʃi /-/pote/) une relation de recouvrement initial (partage de la première syllabe ou des deux premiers phonèmes) avec le mot-cible.

La densité du voisinage phonologique de chaque mot a été estimée à partir de l'outil générateur de voisins phonologiques de la section « Lexique Toolbox », de la base de données Lexique 3.71 (New et al., 2004, cf. Annexe D). Nous avons donc contrôlé ce paramètre, ce dernier pouvant influencer les performances. La densité du voisinage phonologique d'un mot correspondant ici au nombre de mots différant de l'item en question par substitution, addition ou délétion d'un phonème (e.g., Luce & Pisoni, 1998). Par exemple, le mot "poser" présente parmi ses voisins phonologiques « peser » (substitution de phonème), « opposer » (addition de phonème) et « oser » (délétion de phonème). La densité du voisinage phonologique des mots de haute fréquence ($M = 13.9$) ne différait pas statistiquement de celle des mots de basse fréquence ($M = 11.6$), $t(58) = 1.37$, $p > .05$ (cf. Figure 27). Notons que les voisins phonologiques dont la structure syllabique de la première syllabe était différente de l'amorce et de celle de la cible ont été conservés (e.g., mot-cible /po.ze/, « poser » ; voisin /o.poze/, « opposer ») afin de respecter la règle de détermination

du voisinage phonologique au sens de Luce et Pisoni (1998). Remarquons cependant que le fait d'ôter ces voisins n'aurait diminué la taille totale du voisinage total que de 6 % et que cette diminution n'est pas statistiquement significative ($t(110) = 1.22$, $p > .05$).

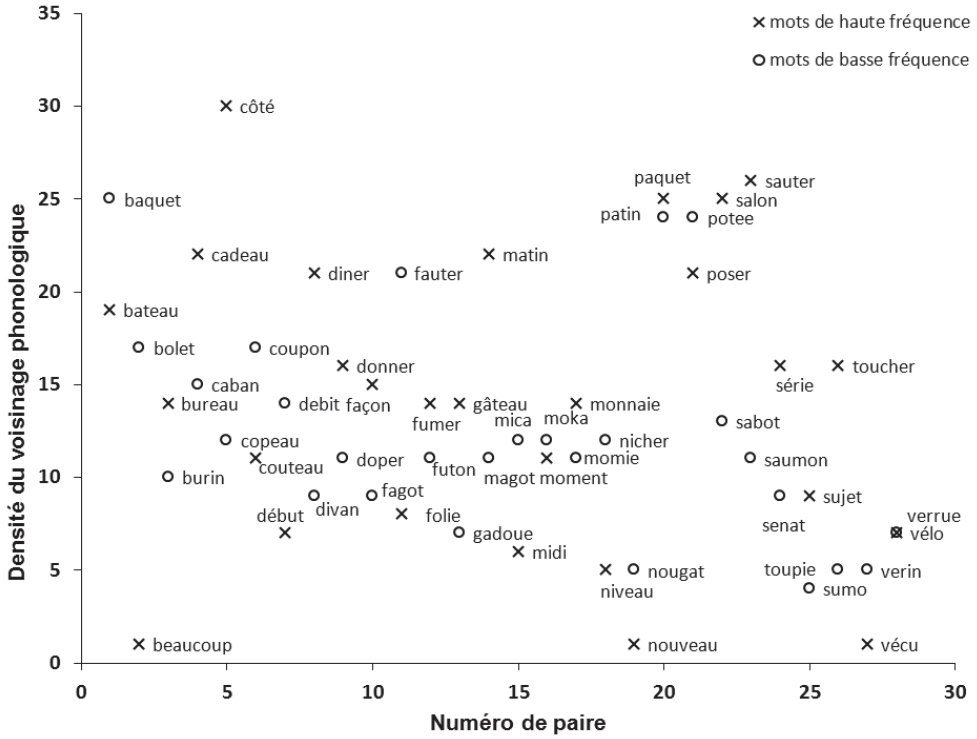


Figure 27. *Densité du voisinage phonologique pour les mots de haute et de basse fréquence de l'Etude 4.*

4.3.1.2.2. Pseudo-mots

De la même manière que dans l'Etude 3, à partir des 60 mots décrits ci-dessus, 60 pseudo-mots bisyllabiques (e.g., /bo.ti/) ont été créés respectant les mêmes caractéristiques de sélection. Ainsi, chacun d'entre eux était également associé avec une amorce reliée (e.g., /bo/-/boti/) et non reliée (e.g., /ʁe /-/boti/). A noter que dans cette étude, aucune analyse ne sera effectuée sur les pseudo-mots, ces derniers permettant uniquement de donner une tâche (i.e., tâche de décision lexicale) à effectuer. Chacun d'entre eux entretenait également un lien de recouvrement initial avec la cible, afin que les participants n'établissent pas leur jugement de lexicalité sur la seule base d'une liaison entre amorces et cibles.

4.3.1.2.3. Items de remplissage

Cent-vingt items bisyllabiques de remplissage (60 mots et 60 pseudo-mots) non reliés avec l'amorce ont également été sélectionnés (e.g., /ʃapo/, « chapeau », /dita/, etc.) réduisant la proportion d'items reliés à 25 % seulement (e.g., Hamburger & Slowiaczek, 1996).

4.3.1.2.4. Enregistrement des stimuli

L'ensemble des stimuli (amorces et cibles) a été enregistré dans une chambre sourde avec le même matériel et dans les mêmes conditions que l'Etude 3. La locutrice était identique à celle de l'Etude 2 et 3. A l'instar de l'Etude 1 et 3, cette dernière était placée devant un fond vert et seulement sa tête, son cou et le haut de ses épaules étaient visibles (cf. Figure 28). La préparation des stimuli a été identique à celle de l'Etude 3. Afin de respecter un ISI de 50 ms entre la présentation de l'amorce et de la cible, le signal vidéo de l'amorce (mais également le signal acoustique de la cible) a été segmenté, ce qui implique que chaque vidéo de l'amorce ne se terminait pas bouche fermée, mais avec un certain degré d'aperture des lèvres[39].

4.3.1.3. Procédure

Les participants étaient placés dans une chambre sourde, à 50 cm environ d'un écran DELL de 16 pouces (1024 x 768 pixels) dont la fréquence de rafraîchissement était fixée à 100 Hertz. Ils entendaient les stimuli par l'intermédiaire d'une enceinte de type SONY SRS-88, située juste derrière l'écran. La procédure était identique à celle utilisée dans l'Etude 1 excepté que l'amorce était toujours présentée en condition V. Afin que la transition amorce/cible ou autrement dit la transition entre les positions lèvres ouvertes de l'amorce/lèvres fermées du visage fixe ne donne pas l'illusion d'un mouvement articulatoire pouvant perturber la concentration des participants, sur chaque première et dernière image de l'amorce la locutrice était présentée avec le bas du visage « flouté » (cf. Figure 28). La dernière image de l'amorce restait à l'écran pendant toute la durée du signal acoustique de la cible. L'expérience se déroulait en un seul bloc ; les participants avaient la possibilité de faire une pause tous les 60 stimuli. A l'intérieur de chaque bloc, le statut lexical des cibles (mot vs. pseudo-mot), la fréquence lexicale des mots-cibles (haute fréquence vs. basse fréquence) ainsi que le type d'amorce (reliée vs. non reliée) étaient présentés dans un ordre aléatoire. Une période d'entraînement de 8 items précédait l'expérience proprement dite. Le déroulement de chaque essai en fonction des conditions est représenté dans la Figure 28. La génération des stimuli et la collecte du type et du temps de réponse était assurée par le logiciel E-Prime 2.0 (*Psychological Software Tools, Pittsburgh, PA*). La totalité de la passation durait 20 minutes environ.

[39] Correspondant à 90 % d'énergie du noyau de la voyelle. A noter que cette mesure a été effectuée sur le signal acoustique de l'amorce, mais que ce dernier n'était jamais joué aux participants, l'amorce étant toujours présentée en modalité visuelle seule.

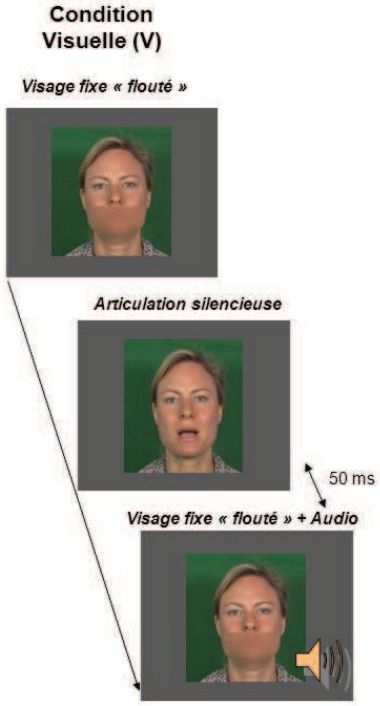


Figure 28. *Représentation schématique du déroulement de l'Etude 4. Pour les items expérimentaux, l'amorce pouvait soit être reliée soit non reliée avec la cible. Lorsque la cible était un mot, celle-ci pouvait être de haute (condition reliée : /po/-/poze/ ; condition non reliée : /ʃi/-/poze/) ou de basse fréquence dans le langage oral (condition reliée : /po/-/pote/ ; condition non reliée : /ʃi/-/pote/).*

4.3.2. Résultats

Le pourcentage d'erreurs ainsi que la moyenne des temps de réponse (mesurés à partir du début du mot-cible, uniquement sur les réponses correctes pour les mots[40]) ont été calculés pour chaque participant et chaque paire d'items. Deux paires d'items (« tirer »- « titan » ; « beauté »- « baudet ») ont été retirés de l'analyse, le membre de basse fréquence de chacune de ces deux paires (i.e., « titan » et « baudet ») ayant un pourcentage d'erreur supérieur à 60 %. Les temps de réponse considérés comme aberrants (i.e., inférieur à 100 ms ou supérieur à 2500 ms) ou se situant à plus de 2 écart-types de la moyenne du participant pour chaque condition respective ont été exclus de l'analyse. Cette opération a écarté 2.5 % des données totales. Une analyse de la variance (ANOVA) 2 (Type d'Amorce : Reliée vs. Non Reliée) x 2 (Fréquence du mot-cible : Haute vs. Basse) a donc été effectuée en échantillon appariés sur 20 participants, par participants (F_1) et par items (F_2), sur les temps de réponse et le pourcentage d'erreurs.

[40] Afin de tester s'il existait un effet d'amorçage pour les pseudo-mots, un test de Student Type d'amorce (Reliée vs. Non Reliée) a été effectué par participants et par items sur les temps de réponses sur les réponses correctes d'une part et les erreurs d'autre part. Aucun effet significatif n'a été mis en évidence.

4.3.2.1. Temps de réponse

Les temps de réponse moyens de chaque condition de l'Etude 4 sont représentés dans le Tableau 11.

Tableau 11. *Temps de réponse (en ms) en fonction des différentes conditions de l'Etude 4. L'erreur type est présentée entre parenthèses.*

Fréquence de la Cible	Type d'amorce		Effet d'amorçage
	Reliée	Non Reliée	
Haute Fréquence	839 (16.1)	852 (16.1)	13
Basse Fréquence	913 (15.6)	974 (17.4)	61

L'analyse sur les temps de réponse a révélé un effet principal du type d'amorce $F_1(1, 19) = 14.33$, $p < .005$, $\eta^2_p = .43$, $F_2(1, 27) = 17.591$, $p < .005$, $\eta^2_p = .39$, mettant en évidence que les participants étaient plus rapides pour reconnaître un mot dans la condition reliée (effet d'amorçage facilitateur). Un effet principal de la fréquence de la cible a également été obtenu, $F_1(1, 19) = 75.15$, $p < .001$, $\eta^2_p = .80$, $F_2(1, 27) = 39.97$, $p < .001$, $\eta^2_p = .60$, indiquant que les participants étaient plus rapides pour reconnaître les mots fréquents que ceux de basse fréquence. Conformément à nos hypothèses, l'effet d'interaction entre ces deux facteurs était également significatif, $F_1(1, 19) = 7.01$, $p < .05$, $\eta^2_p = .27$, $F_2(1, 27) = 12.82$, $p < .005$, $\eta^2_p = .32$. Des comparaisons planifiées ont révélé que l'effet d'amorçage était présent pour les mots de basse fréquence, $F_1(1, 19) = 13.2$, $p < .005$, $\eta^2_p = .59$, $F_2(1, 27) = 27.1$, $p < .001$, $\eta^2_p = .50$, mais pas pour les mots de haute fréquence, $F_1 < 1$.

Des analyses supplémentaires ont été effectuées afin d'évaluer l'impact de la densité du voisinage phonologique sur nos résultats. En effet, la densité du voisinage phonologique est également un paramètre connu pour influencer la perception de la parole à un niveau lexical (e.g., Auer, 2002; Luce & Pisoni, 1998; Vitevitch & Luce, 1998). Ce postulat provient du fait que les modèles actuels (voir McQueen, 2007 pour une revue récente) décrivent le processus de reconnaissance de mots comme résultant d'une activation parallèle de plusieurs candidats lexicaux qui vont entrer en compétition jusqu'à ce que l'un d'eux soit reconnu. Ainsi, plus le nombre de candidats lexicaux activés est important (i.e., lorsque la densité du voisinage phonologique est importante), plus la reconnaissance du mot va prendre du temps et générer des erreurs d'identification. Inversement, plus la compétition lexicale entre les différents candidats est faible, c'est-à-dire moins le mot à identifier dispose de voisins phonologiquement proches et plus sa reconnaissance s'en trouvera facilitée. La densité du

voisinage phonologique constitue donc un paramètre qui, par définition, se situe au niveau lexical, puisque celui-ci influence le processus même de compétition lexicale. Dans cette hypothèse, si la facilitation observée pour la condition visuelle seule de l'Etude 3 résulte de l'intervention de mécanismes lexicaux, nous devrions observer une relation entre la taille de l'effet d'amorçage et la densité de voisinage phonologique du mot-cible. Cette variable a donc été ajoutée en tant que facteur continu dans l'ANOVA par items, pour les temps de réponse.

Les résultats révèlent tout d'abord que l'effet principal de fréquence est conservé, $F_2(1, 27) = 32.3$, $p < .001$, $\eta^2_p = .38$, alors que l'effet principal du type d'amorce n'est plus du tout significatif, $F_2 < 1$. Cependant, les analyses montrent que l'effet d'amorçage interagit toujours avec la fréquence de la cible, $F_2(1, 27) = 13.3$, $p < .001$, $\eta^2_p = .20$, mais également avec la densité du voisinage phonologique, $F_2(1, 27) = 6.1$, $p < .05$, $\eta^2_p = .10$. Afin de mieux caractériser l'influence de la densité du voisinage phonologique sur l'effet d'amorçage, des corrélations ont été effectuées entre ces deux facteurs, pour les mots de haute fréquence d'une part et les mots de basse fréquence d'autre part. Alors que les analyses n'ont indiqué aucune corrélation significative pour les mots de haute fréquence ($r = 0.16$, $p > .05$, cf. Figure 29) ces dernières montrent que pour les mots de basse fréquence la taille de l'effet d'amorçage augmente plus la densité du voisinage phonologique est importante ($r = 0.46$, $p < .05$, cf. Figure 30).

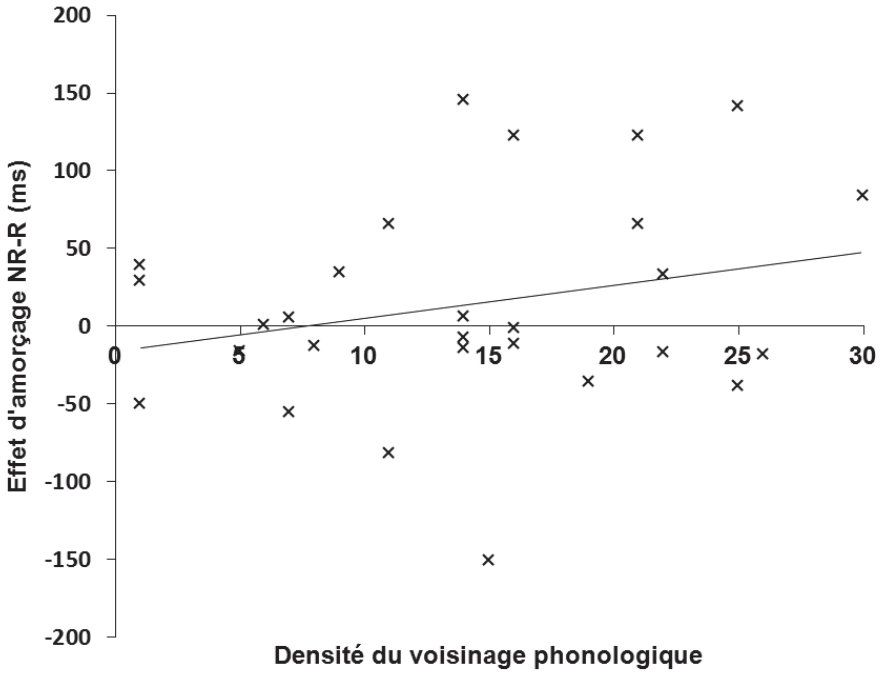


Figure 29. *Effet d'amorçage en fonction de la densité du voisinage phonologique pour les mots de haute fréquence.*

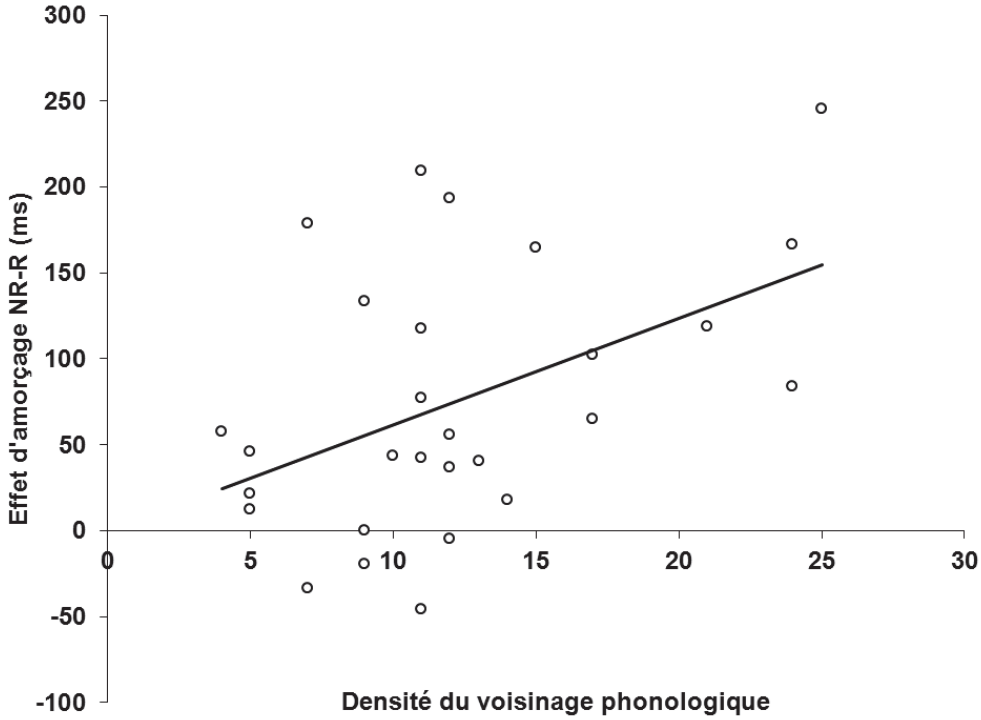


Figure 30. *Effet d'amorçage en fonction de la densité du voisinage phonologique pour les mots de de basse fréquence.*

4.3.2.2. Erreurs

Les pourcentages d'erreurs de chaque condition de l'Etude 4 sont représentés dans le Tableau 12.

Tableau 12. *Pourcentage d'erreurs en fonction des différentes conditions de l'Etude 4. L'erreur type est présentée entre parenthèses.*

Fréquence de la Cible	Type d'amorce		Effet d'amorçage
	Reliée	Non Reliée	
Haute Fréquence	0.3 (0.3)	1.8 (0.9)	1.5
Basse Fréquence	15.6 (4.1)	16.1 (4.6)	1.5

L'analyse sur les pourcentages d'erreurs a révélé un effet principal de la fréquence de la cible, $F_1(1, 19) = 39.21$, $p < .001$, $\eta^2_p = .67$, $F_2(1, 27) = 17.32$, $p < .001$, $\eta^2_p = .39$, mettant en évidence que les participants avaient de meilleures performances pour reconnaître les mots fréquents. Aucun autre effet n'a été obtenu, tous les $F_i < 1$.

Lorsque la densité du voisinage phonologique est ajoutée en tant que facteur continu dans l'analyse par items, l'effet principal de fréquence est conservé, $F_2(1, 27) = 21.7$, $p < .001$,

η^2_p = .29. Un effet principal du voisinage phonologique est significatif, $F_2(1, 27) = 5$, $p < .05$, η^2_p = .08, suggérant que les participants faisaient plus d'erreurs pour reconnaître des mots avec une densité de voisinage importante. Des corrélations ont mis en évidence que cette relation n'était pas significative pour les mots de haute fréquence ($r = 0.13$, $p > .05$) mais seulement pour les mots de basse fréquence ($r = 0.48$, $p < .005$).

4.3.3. Discussion

L'objectif de l'Etude 4 était d'examiner l'impact de la fréquence lexicale du mot-cible sur l'effet d'amorçage obtenu par la présentation d'une amorce en modalité V dans l'Etude 3. Cette manipulation avait pour but de déterminer le locus de la facilitation observée par la présentation d'un geste articulatoire correspondant à la première syllabe d'un mot (e.g., articulation silencieuse de /po/) sur la reconnaissance de celui-ci (e.g., /poze/, « poser »). Pour cela, un paradigme d'amorçage phonologique partiel a été utilisé, l'amorce correspondant à l'articulation silencieuse d'une syllabe, pouvant être reliée (e.g., /po/ → /poze/, vs. /ʃi/ → /poze/) ou non avec la cible, présentée auditivement. Lorsque la cible était un mot, celle-ci pouvait être de haute ou de basse fréquence lexicale (e.g., /poze/ vs. /pote/, respectivement).

Les résultats montrent qu'à l'instar de l'Etude 3, un effet d'amorçage a été obtenu sur les temps de réponse, suggérant que le fait de voir le geste articulatoire de /po/ accélère la reconnaissance auditive de « poser ». Cette facilitation était significative pour les mots de basse fréquence, mais pas pour les mots de haute fréquence (voir e.g., Forster & Davis, 1984, pour des résultats similaires en reconnaissance de mots écrits). Des analyses supplémentaires ont pu mettre en évidence que l'effet d'amorçage observé pour les mots de basse fréquence augmentait plus la densité de voisinage phonologique était importante. Les résultats indiquent également que les participants étaient meilleurs (i.e., étaient plus rapides et faisaient moins d'erreurs) pour identifier des mots de haute que de basse fréquence dans le langage oral. Les analyses sur les erreurs révèlent également que les mots ayant une densité de voisinage phonologique faible étaient mieux reconnus que ceux avec une densité de voisinage élevée.

4.3.3.1. Effet d'amorçage

Les Etudes 3 et 4 ont permis de montrer que le fait de voir l'articulation silencieuse de la première syllabe d'un mot accélère sa reconnaissance auditive ultérieure. Ce résultat

suggère que la présence d'information visuelle facilite la reconnaissance des mots. Les résultats de l'Etude 4 ont indiqué que cet effet de facilitation était modulé par des paramètres tels que la fréquence lexicale ou la densité du voisinage phonologique de la cible. Cela indique que la facilitation observée par la présentation d'une amorce visuelle serait liée à l'intervention de mécanismes lexicaux plutôt que pré-lexicaux. Ces résultats sont en accord avec ceux de Kim et al. (2004, Expérience 1 à 3) et globalement (mais voir section 4.3.3.2.2) avec ceux de Buchwald et al. (2009), puisqu'ils suggèrent que l'information visuelle seule permet d'activer les unités lexicales. Ainsi, par rapport aux travaux de Kim et al. (2004) et de Buchwald et al. (2009), les études présentées dans ce chapitre indiquent que le geste articulatoire *seul* correspondant uniquement aux deux phonèmes initiaux (ou première syllabe) d'un mot apporte suffisamment d'information pour faciliter (i.e., accélérer) sa reconnaissance ultérieure. Cela suggère que l'information visuelle intervient dans les étapes précoces d'activation des différents candidats lexicaux.

Afin d'explorer de manière plus approfondie le rôle de l'information visuelle dans cette étape, nous avons comparé les résultats issus des Etudes 3 et 4 à ceux obtenus dans l'Expérience 4 de Kim et al. (2004). Dans cette étude, les auteurs ont utilisé un paradigme d'amorçage par répétition. A l'instar de l'Expérience 1, les participants percevaient une amorce en modalité visuelle seule (i.e., articulation silencieuse), suivie d'une cible écrite. Les amorces comme les cibles étaient toutes composées de trois phonèmes et pouvaient soit correspondre toutes les deux à un mot, soit toutes les deux à un pseudo-mot. Dans une première condition, l'amorce partageait le même début avec la cible (e.g., pour les mots : /bak/, « back » → /bænd/, « BAND », groupe ; pour les pseudo-mots : /skeɪ/, « scay » → /skap/, « SCAP »). Dans une seconde condition l'amorce partageait la même fin (rime) avec la cible (e.g., pour les mots : « back » → /sak/, « SACK », sac ; pour les pseudo-mots : « scay » → /veɪ/, « VAY »). Dans la condition contrôle, l'amorce ne partageait aucun lien phonologique avec la cible (e.g., pour les mots : /liːf/, « leaf » → « BAND » ou « SACK » ; pour les pseudo-mots : /twɪʃ/, « twish » → « SCAP » ou « VAY »). Dans cette expérience, les auteurs ont mis en évidence un effet d'inhibition de la présentation de l'amorce lorsque celle-ci partageait le même début (mais pas la même fin) que la cible par rapport à la condition contrôle, uniquement pour la condition mot. Aucun effet de facilitation ou d'inhibition n'a été mise en évidence pour la condition rime par rapport à la condition contrôle, mettant en avant l'importance du début de mot (par rapport à la fin) dans le processus d'accès au lexique (voir e.g., Spinelli, 1999 ; Spinelli et al., 2001, pour des résultats similaires en modalités auditive). Ainsi, leurs résultats montrent que l'articulation silencieuse

du mot « back » inhibe la représentation du mot « band » (Expérience 4). Cet effet est couramment retrouvé dans la littérature étudiant l'accès au lexique en modalité auditive (e.g., Dufour & Peereman, 2003). Cela signifie que le geste articulatoire seul, correspondant à la fin de l'amorce (e.g., /ak/ de /bak/, « back ») fournit une information visuelle suffisante pour inhiber la représentation lexicale du mot-cible (e.g., « BAND »). Or, les Etudes 3 et 4 de ce chapitre montrent que l'articulation silencieuse correspondant aux deux phonèmes initiaux d'un mot permet également d'en activer la représentation lexicale.

En conclusion, ces données indiquent que de la même manière que l'information auditive (cf. Spinelli et al., 2001), l'information visuelle relative au début de mot joue un rôle prépondérant dans l'accès au lexique. Ensuite, ces résultats permettent de montrer que l'information visuelle permet soit d'inhiber soit de faciliter la reconnaissance ultérieure d'un mot-cible. Cela suggère que l'information visuelle seule permet tout aussi bien d'augmenter mais également de *réduire* le nombre de candidats lexicaux, lors du processus de compétition lexicale.

Dans la vie quotidienne cependant, l'information auditive est rarement complètement absente des interactions langagières, rendant les situations de perception de la parole en modalité visuelle seule relativement rares. Il paraît donc légitime de se demander si l'information visuelle permet de jouer un rôle dans l'accès au lexique en présence de l'information auditive. Dans le Chapitre 3, nos données ont montré que le fait de voir les gestes articulatoires de son interlocuteur permet d'accéder plus efficacement (Etude 1 et 2) mais également plus rapidement (Etude 2) aux unités lexicales, en présence d'une information auditive détériorée. Bien que cette idée dépasse les conclusions que nous pouvons tirer des Etudes 3 et 4, il serait envisageable que l'information visuelle seule permette de pré-activer un certain nombre de candidats lexicaux *avant* que l'information auditive ne soit disponible au niveau lexical. Ce prétraitement permettrait ainsi d'*anticiper* l'arrivée de l'information auditive et ainsi d'accélérer (selon le modèle de la Cohorte II) la formation de la cohorte initiale (voir e.g., Marslen-Wilson & Warren, 1994 pour une idée similaire en modalité auditive). Rappelons qu'en accord avec cette idée d'anticipation, des travaux conduits par Cathiard (1994) et Smeele (1994) ont montré que notre système visuel est capable de décoder l'information visuelle (correspondant ici aux mouvements des lèvres d'un locuteur) afin d'identifier un phonème, avant même que toute information auditive relative à ce dernier ne soit disponible dans le signal acoustique. Ces études suggèrent donc que l'information visuelle (i.e., l'arrondissement des lèvres, Cathiard, 1994 ; l'occlusion

labiale, Smeele, 1994) serait traitée par notre système perceptif afin d'*anticiper* sa conséquence acoustique (voir Van Wassenhove, et al., 2005, pour des résultats en neuro-imagerie compatibles avec cette hypothèse). Les résultats de l'Etude 2 (cf. Chapitre 3) pourraient également être en accord avec cette hypothèse. Il serait donc envisageable que ce traitement permette à l'information visuelle d'activer les unités pré-lexicales avant que l'information acoustique ne soit disponible à ce niveau. Une cohorte initiale « visémique » dont l'ensemble des candidats lexicaux correspondrait au(x) caractéristique(s) articulatoire(s) visible(s) du signal visuel d'entrée (e.g., place d'articulation bilabiale ; occlusive) pourrait être générée (e.g., « poser »; « peser », etc.) avant l'arrivée de l'information auditive. Cette hypothèse sera plus largement développée dans la section 6.2.2.1 du Chapitre 6.

4.3.3.2. Influence de la fréquence et du voisinage phonologique de la cible

Les Etudes 3 et 4 mettent en évidence que la reconnaissance d'un mot est facilitée lorsqu'elle est précédée de la présentation d'un geste articulatoire correspondant à la première syllabe de ce même mot. Cet effet d'amorçage a été majoritairement constaté pour des mots peu fréquents dans le langage oral (Etude 4). De plus, indépendamment de cette influence, nous avons également observé que la taille de cette facilitation augmentait plus la densité du voisinage phonologique était importante. De par l'influence de ces deux paramètres sur cet effet d'amorçage, ces résultats indiquent que les mécanismes responsables de cette facilitation sont plutôt de nature lexicale plutôt que pré-lexicale. Les deux paragraphes suivants discutent de l'influence respective de la fréquence d'occurrence et de la densité du voisinage phonologique sur l'ensemble de nos résultats.

4.3.3.2.1. Influence de la fréquence

Dans l'Etude 4, le mot-cible était exclusivement présenté en modalité auditive. Conformément aux résultats de la littérature, nous avons observé que les participants étaient plus rapides et faisaient moins d'erreurs pour reconnaître un mot de haute fréquence plutôt qu'un mot de basse fréquence. Cette différence a été obtenue alors que des paramètres tels que la durée ou le point d'unicité phonologique des cibles entre les mots fréquents et les mots peu fréquents aient été contrôlés (cf. Goldinger, 1996 ; Marslen-Wilson, 1990). En conséquence, nos données indiquent que, conformément aux résultats de la littérature (e.g., Taft & Hambly, 1986), les mots fréquents étaient plus facilement reconnus que les mots de basse fréquence lexicale.

Nous avons également observé que la fréquence des mots-cibles avait un impact sur la taille de l'effet d'amorçage obtenu. Ce résultat suggère que l'effet de facilitation observé par la présentation d'une amorce en modalité V se situe au niveau lexical. En d'autres termes, le fait de voir le geste articulatoire pour /po/ permettrait d'accélérer la reconnaissance de /pote/, « potée » parce que l'articulation silencieuse pour la première syllabe d'un mot apporte suffisamment d'information visuelle pour activer sa représentation lexicale. L'information visuelle interviendrait donc dans les phases précoces de l'accès au lexique. Or, dans l'Etude 4, l'effet de facilitation n'a été observé que pour les mots-cibles de basse fréquence. Le sens de cette interaction est en accord avec les données issues de la littérature portant sur la reconnaissance de mots écrits (e.g., Forster & Davis, 1984). Ce résultat peut être expliqué par le fait qu'un mot de haute fréquence n'a besoin que de peu d'activation pour être reconnu. Ainsi, la facilitation fournie par l'information visuelle seule étant relativement faible (cf. section 4.2.3), celle-ci ne serait notable que pour les mots de basse fréquence, les performances pour reconnaître les mots de haute fréquence étant au plafond (M_{erreur} = 1.5 %). Ce résultat suggère donc que voir le visage de son interlocuteur contribue au processus de perception de la parole lorsque l'accès au lexique est difficile. Remarquons que cette idée a déjà été évoquée pour la reconnaissance de mots en modalité audiovisuelle au chapitre précédent (Chapitre 3). Notons néanmoins que les travaux de Buchwald et collègues (2009) semblent obtenir des résultats inverses aux nôtres, c'est-à-dire un plus grand bénéfice de la présence d'une amorce en modalité V pour la reconnaissance des mots fréquents dans le langage oral. Autrement dit, ces auteurs suggèrent que la présence de l'information visuelle favoriserait principalement le processus de reconnaissance d'un mot présenté auditivement lorsque celui-ci est facile à identifier. Ce point sera plus largement débattu dans la section suivante.

4.3.3.2.2. Influence du voisinage phonologique

Plusieurs études ont mis en évidence que des mots ayant une faible densité de voisinage phonologique (i.e., ayant peu de voisins) étaient mieux reconnus que des mots ayant une densité de voisinage importante. Ce résultat a été trouvé pour des mots présentés en modalité auditive seule (e.g., Buchwald, et al., 2009; Goldinger, et al., 1989), audiovisuelle (e.g., Kaiser et al., 2003) et visuelle seule (e.g., Auer, 2002). Sur ce point, les résultats de l'Etude 4 sont en accord avec les données de la littérature. En effet, nous avons observé que plus la densité du voisinage phonologique du mot-cible auditif augmentait, plus celui-ci engendre des erreurs d'identification. Ce résultat est majoritairement interprété comme le fait que plus un mot a de voisins phonologiquement similaires, plus le nombre de candidats

lexicaux activés lors de la reconnaissance de ce mot est important, plus la reconnaissance du mot va générer des erreurs d'identification. Inversement, plus la compétition lexicale entre les différents candidats est faible, c'est-à-dire moins le mot à identifier dispose de voisins phonologiquement proches et plus sa reconnaissance s'en trouvera facilitée.

Or, nous avons également observé que lorsque l'effet d'amorçage (obtenu sur les temps de réponse) était significatif (e.g., pour les mots de basse fréquence) celui-ci était également corrélé positivement avec la densité du voisinage phonologique. Le fait que la densité du voisinage phonologique comme la fréquence du mot-cible interagissent avec l'effet d'amorçage indique qu'il existe un impact spécifique de chacun de ces facteurs sur cette facilitation. Ce résultat vient corroborer le fait que la facilitation observée est bien due à l'intervention de mécanismes lexicaux plutôt que pré ou post-lexicaux. En effet, cette donnée indique que l'impact de l'information visuelle dans le processus de reconnaissance de mots serait fonction du nombre de candidats lexicaux activés lors de la présentation du mot-cible, en modalité auditive. Le sens de ces corrélations indique que plus la densité du voisinage phonologique augmente, plus l'effet de facilitation est plus important. Ce résultat pourrait paraître en opposition avec celui observé par Buchwald et al. (2009). En effet, leurs données indiquent que l'information visuelle seule facilite principalement la reconnaissance de mots fréquents dans le langage oral, ayant une densité de voisinage phonologique faible. Cependant, de nombreuses différences méthodologiques entre ce travail et l'Etude 4 pourraient expliquer les différences de résultats. Ces auteurs ont utilisé un paradigme d'amorçage par répétition consistant à présenter l'articulation silencieuse d'un mot *entier* (condition reliée) ou le *visage immobile du locuteur* (condition contrôle) comme amorce suivie de la cible, après un ISI de *500 ms*. Ils ont observé un effet d'amorçage plus important pour les mots fréquents et ayant peu de voisins en calculant *le nombre de segments de la cible correctement identifiés*, celle-ci étant présentée en modalité auditive avec un signal acoustique *détérioré* (i.e., présence de bruit). Or, dans l'Etude 4, nous avons utilisé un paradigme d'amorçage phonologique partiel où nous avons présenté le geste articulatoire correspondant soit à la première syllabe (ou aux deux phonèmes *initiaux*) de la cible (condition reliée) soit à deux *phonèmes différents* (condition non reliée) en tant qu'amorce. Cette amorce était suive par la cible présentée en modalité auditive *sans bruit*, après un ISI de *50 ms*. Nos résultats sont observés sur les *temps de réponse* à l'aide d'une *tâche de décision lexicale* effectuée sur la cible. Nous pensons que de par l'utilisation de stimuli et de critères méthodologiques divergents, l'étude de Buchwald et collègues évalue des processus différents de la nôtre. En effet, leur paradigme interroge le processus de reconnaissance de mots avec une mesure effectuée en différé par rapport à la tâche (i.e., lorsque le processus de reconnaissance de mots est achevé).

Au contraire, notre paradigme permet d'interroger les phases précoces de l'accès au lexique avec une mesure réalisée en temps réel, lors du déroulement de la tâche (i.e., temps de réponse). Notons que cette remarque est également valable pour les travaux ayant étudié l'impact de la densité de voisinage phonologique et de la fréquence lexicale sur l'identification de mots en modalité visuelle seule (e.g., Kaiser et al., 2003 ; Tye-Murray et al., 2009 ; Auer, 2002 ; Mattys et al., 2002 ; Feld et al., 2011).

Ainsi, nous pensons que l'information visuelle jouerait un rôle spécifique dans les phases précoces de l'accès au lexique et ce majoritairement lorsque le mot est difficile à reconnaître (i.e., voisinage dense et mots peu fréquents).

4.4. CONCLUSIONS

L'objectif de ce chapitre était d'explorer le rôle de l'information visuelle dans les phases précoces de l'accès au lexique chez l'adulte. Les résultats montrent que le fait de voir le geste articulatoire correspondant à la première syllabe (ou aux deux phonèmes initiaux) d'un mot constitue une information suffisante pour activer sa représentation lexicale. L'information visuelle *seule* jouerait un rôle prépondérant dans les phases *précoces* du processus de reconnaissance de mots lorsque l'accès au lexique est difficile (mots de basse fréquence, ayant une densité de voisinage phonologique élevée). Ces données témoignent donc de la nécessité de considérer le processus de reconnaissance de mots comme un évènement bimodal (voire même multimodal, cf. Chapitre 1). Cependant la majorité des modèles actuels décrivant l'accès au lexique (e.g., Merge, TRACE, Cohorte II, etc.) ne considère pas l'information visuelle comme source d'information. Ce travail constitue donc une preuve supplémentaire selon laquelle ces derniers devraient incorporer la composante articulatoire du signal parole dans leur architecture. Nous essaierons d'envisager comment le signal visuel de parole pourrait être inclus dans ces modèles dans la section 6.1.3 du Chapitre 6. Avant cela, nous allons étudier dans le prochain chapitre si l'information visuelle participe également au processus d'activation des représentations lexicales chez l'enfant.

CHAPITRE 5. APPORT DE L'INFORMATION VISUELLE ET LEXICALE A LA PERCEPTION DE LA PAROLE CHEZ L'ENFANT

« Perhaps the most salient sight to the newborn is the human face. […] Faces are a major instrument of affective and social communication, and I believe faces play a developmental linguistic role : they help launch the normal neonate on a trajectory that, with other capabilities and experiences, will lead to the acquisition of spoken language. For the face not only conveys emotion directly, it is also the origin of the human voice."

John Locke (1993). p 42.

5.1. INTRODUCTION

Chez l'adulte, plusieurs études ont mis en évidence que le fait de voir le visage de son interlocuteur en mouvement permet d'augmenter l'intelligibilité du signal de parole lorsque l'information auditive est détériorée (Benoît et al., 1994 ; Binnie et al., 1974 ; Erber, 1969 ; Sumby & Pollack, 1954 ; cf. Chapitre 1). Par exemple, Benoît et al. (1994) ont montré que lorsque le signal acoustique était accompagné de bruit blanc, des phonèmes consonantiques et vocaliques présentés dans des non-mots étaient mieux identifiés en modalité audiovisuelle qu'auditive seule. En conséquence, décoder les mouvements articulatoires du visage de son interlocuteur permet d'augmenter l'intelligibilité des phonèmes (décodage pré-lexical), lorsque l'information auditive est dégradée. Les résultats issus des Chapitres 3 et 4 indiquent que le système visuel de l'adulte code les gestes articulatoires de parole afin d'en extraire des informations qui sont exploitées durant le processus de reconnaissance des mots, c'est-à-dire lors de l'activation des représentations lexicales.

Cette question n'a, à notre connaissance, jamais été explorée chez l'enfant. Ainsi, l'objectif de la présente recherche est d'étudier si cette capacité à décoder visuellement les gestes de parole et à utiliser cette information dans le processus d'accès au lexique est présente durant l'enfance. Nous allons donc passer en revue différents travaux qui ont cherché à évaluer l'apport de l'information visuelle à la perception de la parole chez l'enfant et le nourrisson. Nous examinerons ensuite plusieurs études qui se sont intéressées au processus de reconnaissance de mots en modalité auditive chez ces mêmes populations.

5.1.1. Apport de l'information visuelle à la perception de la parole chez l'enfant et le nourrisson

5.1.1.1. Etudes chez le nourrisson

Un grand nombre d'études ont pu montrer que l'être humain est, depuis les premières phases de son développement, capable de traiter le signal visuel de parole (Burnham & Dodd, 2004; Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2002, 2003; Rosenblum, Schmuckler, & Johnson, 1997; Weikum, et al., 2007), (voir Burnham, 1998; Soto-Faraco, et al., sous presse, pour une revue récente; Woodhouse, Hickson, & Dodd, 2009). Patterson et Werker (2003) ont par exemple étudié la capacité de nourrissons à détecter la correspondance entre le signal acoustique et visuel de parole. Pour cela, ces auteurs ont, dans une première phase de l'expérimentation, exposé des nourrissons âgés de 2 mois à des

voyelles isolées (phase de familiarisation). Lors de cette phase, ces voyelles (i.e., /a/ et /i/) étaient premièrement présentées pendant 9 secondes chacune en modalité visuelle seule (articulation silencieuse du locuteur sans signal auditif) sur un premier puis un second écran, tous deux situés à égale distance de l'enfant. La même procédure était ensuite répétée pendant neuf secondes supplémentaires, excepté que le signal acoustique des stimuli accompagnait également l'articulation de chacune des voyelles correspondante. Lors de la phase de test, chaque stimulus visuel était présenté simultanément : sur un des deux écrans le visage du locuteur articulait un /i/, alors que sur l'autre, le visage du locuteur articulait un /a/. Le signal acoustique joué pendant cette phase pouvait correspondre soit à un /a/, soit à un /i/. Leurs résultats montrent que lors de cette phase les nourrissons regardaient plus longtemps le visage du locuteur dont le geste articulatoire était congruent avec l'information auditive. Ainsi, ces résultats indiquent que les nourrissons seraient capables, dès l'âge de 2 mois, de détecter la correspondance entre la composante auditive et visuelle du signal de parole (voir Burnham, 1993, pour une correspondance visage-voix dès l'âge de 1 mois).

L'étude de Burnham et Dodd (2004) suggère même que des nourrissons âgés de 4 mois et demi possèderaient la capacité à intégrer ces deux sources d'information. Dans ce travail, le paradigme de l'effet McGurk (McGurk & MacDonald, 1976) a été adapté à l'étude des compétences précoces des nourrissons. Rappelons que cette illusion perceptive, la perception d'un /da/ ou /ða/, émerge chez l'adulte lors de la présentation acoustique d'un /ba/ doublée de l'articulation d'un /ga/. Celle-ci indique que les informations visuelles sont intégrées à l'information auditive. Dans leur étude, Burnham et Dodd (2004) ont utilisé cet effet avec un paradigme d'habituation-réaction à la nouveauté. Le paradigme d'habituation-réaction à la nouveauté est une mesure indirecte de la perception des nourrissons. Cette technique se déroule selon deux phases : (1) une phase d' «habituation » pendant laquelle, un ou plusieurs stimuli sont présentés à des bébés jusqu'à ce que le temps d'exploration perceptive diminue ; (2) une phase de « test », où l'on présente soit les mêmes stimuli que dans la phase d'habituation, soit des stimuli nouveaux. Dans cette dernière situation, on s'attend à une augmentation des temps d'exploration, suggérant que le bébé est capable de discriminer les stimuli présentés dans les deux phases. Dans cette expérience, lors de la phase d'habituation, la moitié des nourrissons (groupe expérimental) était habituée à un stimulus audiovisuel de type McGurk (i.e., /ba/ auditif + /ga/ visuel) alors que l'autre moitié (groupe contrôle) était habituée à un stimulus audiovisuel congruent (i.e., /ba/ auditif + /ba/ visuel). Lors de la seconde phase, trois stimuli (i.e., /ba/, /da/ ou /ða/) étaient présentés en modalité auditive seule. L'hypothèse des auteurs était que si les nourrissons

sont capables d'intégrer l'information auditive et visuelle, le percept auquel le groupe expérimental devrait être familiarisé devrait correspondre à /da/ ou /ða/. Dans le cas contraire, le percept avec lequel ce groupe est le plus familier devrait être identique à celui du groupe contrôle, c'est-à-dire /ba/ (présenté en modalité auditive). Leurs résultats indiquent que les stimuli /da/ et /ða/ étaient considérés comme familiers par le groupe expérimental, mais pas par le groupe contrôle. De ce fait, cette différence suggère que les nourrissons du groupe expérimental percevaient plutôt /da/ ou /ða/ dans la phase d'habituation, autrement dit qu'ils étaient sensibles à l'illusion McGurk (voir Rosenblum et al., 1997 pour des résultats similaires). Ainsi, cette étude indique que des nourrissons âgés de 4 mois et demi pourraient non seulement décoder l'information visuelle mais possèderaient également la capacité d'intégrer cette information au signal acoustique de parole.

Weikum et al. (2007) ont, quant à eux, étudié les compétences de décodage du signal visuel seul de parole chez des nourrissons âgés de 4, 6 et 8 mois. L'objectif était d'explorer leur capacité à discriminer deux langues relativement proches (anglais vs. français), sur la seule base de l'information visuelle (articulation silencieuse de phrases, voir Soto-Faraco et al., 2007, pour des résultats chez l'adulte). Les compétences des nourrissons ont été évaluées à l'aide d'un paradigme d'habituation-réaction à la nouveauté. La phase d'habituation consistait à présenter trois locuteurs différents bilingues français-anglais prononçant des phrases soit en français, soit en anglais. Chaque essai correspondait à un passage différent extrait d'une histoire pour enfants. Notons qu'aucune information auditive ne leur était présentée, seul le visage des locuteurs en mouvement. Cette présentation était effectuée jusqu'à ce que le temps de regard de l'enfant diminue de 60 % pendant trois essais consécutifs par rapport au temps de regard mesuré sur les trois premiers essais. Dans la phase de test, la moitié des nourrissons se voyait présenter les mêmes extraits articulés par les mêmes locuteurs mais dans une langue différente par rapport à la phase d'habituation (groupe expérimental). L'autre moitié des enfants se voyait présenter des extraits différents, mais articulés par des locuteurs identiques et dans la même langue que précédemment. Leurs résultats ont mis en évidence que les nourrissons âgés de 4 mois montraient un regain d'intérêt lors de la phase de test (i.e., augmentation du temps de regard par rapport à la phase d'habituation) lors du changement de langue et non lors du changement d'extrait. Cette première donnée suggère donc que dès l'âge de 4 mois, un enfant dispose de capacités d'analyse de l'information visuelle suffisamment fines pour discriminer le français de l'anglais sur la seule base du geste articulatoire. A 6 et 8 mois, les données ont montré que seuls les nourrissons bilingues anglais-français conservaient cette capacité à distinguer ces

deux langues. Ces résultats indiquent que nous disposons de compétences précoces pour traiter les mouvements du visage de nos interlocuteurs mais aussi que ces compétences se spécialiseraient en fonction de notre environnement linguistique.

En conclusion, l'ensemble de ces travaux suggère que dès les premières phases de vie (i.e., de 1 à 24 mois), un nourrisson est capable de décoder et exploiter des informations visuelles du geste de parole. L'objectif de la partie suivante consiste à présenter les différentes études qui se sont intéressées à l'évolution de cette compétence au cours de l'enfance.

5.1.1.2. Etudes chez l'enfant

Bien que les nourrissons semblent être sensibles à l'information visuelle dans leur perception du signal de parole, cette sensibilité ne semble pas être aussi claire à des stades plus avancés de l'enfance, plus particulièrement entre 2 et 14 ans. Ainsi, dans leur étude princeps, McGurk et MacDonald (1976) ont étudié l'effet McGurk chez des participants adultes mais aussi chez des enfants âgés de 3-5 ans et de 7-8 ans. Ces auteurs ont montré que le pourcentage de réponses correspondant à une intégration des informations visuelles et auditives était plus faible pour l'ensemble des enfants par rapport aux adultes. Quelques années plus tard, Massaro et collègues (Massaro, Thompson, Barron, & Laren, 1986) ont trouvé cette même différence de prise en compte de l'information visuelle entre enfants et adultes. Dans cette étude, les expérimentateurs ont demandé à des adultes et des enfants âgés entre 4 et 6 ans d'identifier un signal de parole correspondant au signal acoustique /ba/ doublé de l'articulation d'un /da/. Les performances des enfants montrent que leurs réponses étaient moins dominées par le signal visuel (i.e., moins de réponses /da/) que celles des adultes. Hockley et Polka (Hockley & Polka, 1994) ont également trouvé des résultats indiquant que l'influence de la modalité visuelle sur la perception de la parole augmentait en fonction des différents stades de développement. Ces derniers ont étudié l'effet McGurk chez des participants adultes et des enfants âgés de 5, 7, 9 et 11 ans, en combinant la syllabe auditive /ba/, avec le geste articulatoire /va/, /ða/, /da/, ou /ga/. Leurs données indiquent que l'influence de l'information auditive sur la perception de ces syllabes diminuait avec l'âge, alors que celle de l'information visuelle augmentait (voir Dupont, Aubin, & Ménard, 2005; Tremblay et al., 2007, pour des résultats similaires). De plus, ces derniers ont mis en évidence que seulement la moitié des enfants les plus âgés (i.e., les 10-12 ans) avaient des

performances comparables à celles des adultes, suggérant que la capacité à utiliser le signal visuel de parole continue à se développer après l'âge de 12 ans.

Un travail relativement récent a permis d'étudier le bénéfice lié à la présence d'une information visuelle sans utiliser l'effet McGurk, en présentant un signal acoustique de parole détérioré par du bruit (RSB = -4 dB), chez des enfants de langue maternelle anglaise et japonaise, âgés de 6, 8 et 11 ans (Sekiyama & Burnham, 2008). Classiquement, chez l'adulte, en situation bruitée, de meilleures performances sont observées en modalité audiovisuelle par rapport à une situation auditive seule (Sumby & Pollack, 1954). Dans cette étude, en utilisant une tâche d'identification de syllabe, les auteurs ont montré qu'à l'âge de 6 ans, l'apport de la gestualité articulatoire était faible quelle que soit la langue d'appartenance des enfants, ce qui n'était pas le cas pour les enfants plus âgés. Ce pattern indique donc, conformément aux études citées précédemment, que la capacité à utiliser l'information visuelle (ici, pour augmenter l'intelligibilité du signal acoustique) augmenterait avec l'âge et/ou l'expérience. Cependant une observation plus détaillée des résultats leur a permis de mettre en évidence que le bénéfice lié à la présence de l'information visuelle augmentait majoritairement pour les enfants anglophones (entre 6 et 8 ans), mais restait stable pour les japonais. Pour ce dernier groupe, l'impact de l'âge sur cette compétence s'est révélé très faible, voire inexistant. Ainsi, cette étude suggère que la prise en compte du signal visuel de parole est différente en fonction de l'âge et/ou de l'expérience des participants mais également qu'elle varie selon la langue.

Un travail relativement récent (Jerger, Damian, Spence, Tye-Murray, & Abdi, 2009) s'est intéressé aux capacités de décodage de l'information visuelle à différents stades de l'enfance. A l'instar de l'étude de Sekiyama et Burnham (2008), ces auteurs n'ont pas utilisé l'effet McGurk. En effet, dans leur étude, Jerger et al. (2009) ont employé le paradigme du « multimodal picture-word naming task[41] ». Celui-ci avait pour objectif d'évaluer les compétences perceptives d'enfants âgés de 4 à 14 ans, en utilisant une mesure en temps réel des performances (i.e., temps de réponse). L'hypothèse sous-jacente au paradigme originel du « picture-word naming task » (Jerger, Martin, & Damian, 2002) est que la présentation simultanée d'un distracteur auditif facilite la dénomination d'une image dont le nom commence par le même phonème (e.g., distracteur auditif relié « peach », /piːtʃ/ à l'image « pizza », /piːtza/) par rapport à un distracteur ne partageant aucun phonème avec la cible (e.g., distracteur auditif non relié « eagle », /iːgʌl/ à l'image « pizza »). Dans leur étude, Jerger et al. (2009) demandaient aux enfants de dénommer une image qui était située sur la

[41] Littéralement : tâche de dénomination d'image-mot multimodale

poitrine du locuteur. Le terme « multimodal », vient du fait que le distracteur pouvait soit être présenté en modalité auditive seule, soit en modalité audiovisuelle. Dans la condition audiovisuelle, les enfants pouvaient entendre mais également voir le visage du locuteur lors de l'articulation du mot distracteur. Leurs résultats montrent que les participants étaient plus rapides pour dénommer une image lorsque celle-ci était accompagnée d'un distracteur relié plutôt que non relié, alors même qu'il leur était demandé d'ignorer ce dernier. Les auteurs ont également mis en évidence que cet effet de facilitation était plus important lorsque le distracteur était présenté en modalité audiovisuelle plutôt qu'auditive seule. Cependant, ce bénéfice n'a été observé que pour les groupes d'enfants les plus jeunes (4 ans) et les plus âgés (10-14 ans), mais pas pour ceux ayant un âge intermédiaire (5, 6-7 et 8-9 ans). Autrement dit, bien que les participants âgés de 5 à 9 ans aient bénéficié de la présentation du distracteur relié avec la cible, aucun effet de facilitation supplémentaire n'était observé lorsque le visage en mouvement du locuteur était visible. Afin de s'assurer que cette absence de facilitation n'était pas due à une incapacité à traiter l'information visuelle pour ces groupes d'âge (voir Massaro, 1984; Massaro, et al., 1986, pour une telle interprétation), une tâche de lecture labiale a également été administrée à l'ensemble des enfants (« Children Audiovisual Enhancement Test », CAVET, Tye-Murray & Geers, 2001). Leurs résultats montrent que les scores de lecture labiale augmentaient avec l'âge. Cette donnée réfute ainsi l'hypothèse que l'absence de facilitation supplémentaire en modalité audiovisuelle pour les enfants de 5 à 9 ans puisse être due à une compétence en lecture labiale plus faible que chez les 4 ans. Cette absence d'effet chez les enfants âgés de 5 à 9 ans pourrait être due à une diminution de la *sensibilité* à l'information visuelle. Les auteurs suggèrent que cette baisse de sensibilité serait due à une réorganisation des représentations phonologiques à ce stade de développement. En effet, pendant cette période, les enfants anglais apprennent à lire et développent par conséquent leur conscience phonémique[42]. Dans cette étude, les auteurs argumentent que lors de la tâche de dénomination, la majorité des ressources cognitives des enfants de 5 à 9 ans serait mobilisée pour traiter le signal sonore des distracteurs en tant que phonèmes et qu'ils ne disposeraient pas de ressources supplémentaires suffisantes afin de traiter efficacement l'information visuelle. Les enfants ne disposant pas de conscience phonémique (i.e., 4 ans) ou maîtrisant cette compétence (i.e., 10-14 ans) auraient les ressources cognitives suffisantes pour décoder les gestes articulatoires

[42] Cette habileté désigne la capacité d'identifier, conscientiser, segmenter et manipuler les phonèmes de la langue parlée. Cette compétence est nécessaire à l'acquisition de la lecture, puisque cette activité consiste à associer un phonème à une lettre ou groupe de lettres (graphème).

du locuteur et bénéficier de cette information. Cette hypothèse, bien que plausible, n'a pas été testée par les auteurs.

En conclusion, un certain nombre d'études indique que cette capacité à décoder le signal visuel de parole serait plus faible chez les enfants que chez les adultes (Hockley & Polka, 1994; Massaro, 1984; Massaro, et al., 1986; McGurk & MacDonald, 1976; Sekiyama & Burnham, 2008) et que celle-ci augmenterait avec l'âge et/ou l'expérience (Dupont, et al., 2005; Hockley & Polka, 1994; Sekiyama & Burnham, 2008; Tremblay, et al., 2007). Néanmoins, les mécanismes impliqués dans le développement de cette capacité sont toujours mal compris (Jerger, et al., 2009; Sekiyama & Burnham, 2008) et aucune recherche n'a, à notre connaissance, étudié la contribution spécifique de l'information visuelle dans le processus de reconnaissance de mots chez l'enfant.

5.1.2. Le processus de reconnaissance de mots chez le nourrisson et l'enfant

La réalisation d'un signal de parole peut varier selon un grand nombre de facteurs inter et intra-individuels tels que le genre et l'âge du locuteur, le débit de parole, la présence de bruit environnant, etc. (cf. Chapitre 2). Reconnaître un mot parlé relève donc d'une opération complexe, qui peut s'avérer particulièrement ardue pour les jeunes apprenants d'une langue. Pour reconnaître un mot, l'enfant doit faire abstraction de toute cette variabilité et se concentrer sur les paramètres qui vont lui permettre d'identifier le mot prononcé. Les informations visuelles sur l'articulation du mot pourraient donc être exploitées par les enfants pour faciliter le processus de reconnaissance des mots. Dans cette section, nous allons présenter plusieurs études ayant étudié les mécanismes impliqués dans le processus de reconnaissance des mots chez le nourrisson et l'enfant.

5.1.2.1. Construction des représentations lexicales chez le nourrisson et le jeune enfant

Afin de pouvoir attribuer aux différentes occurrences d'un même mot la même signification, il est communément admis que chaque être humain adulte possède un lexique mental (Treisman, 1960, cité dans Spinelli & Ferrand, 2005). Ce dernier regrouperait une ou plusieurs représentations en mémoire des mots connus par un individu (cf. Chapitre 2). Chaque représentation dite « lexicale » disposerait de caractéristiques spécifiques (e.g.,

phonémiques, phonologiques, etc.) permettant d'identifier un mot en dépit des variations et de la continuité du signal de parole (cf. Chapitre 2). L'âge à partir duquel un enfant dispose de telles représentations a ainsi largement été questionné dans la littérature.

Pour cela, certains travaux ont montré que les enfants ont, dès l'âge de 7.5 mois, la capacité de reconnaître un mot inséré dans une phrase lorsque celui-ci a été présenté isolément auparavant (Jusczyk & Aslin, 1995). Dans cette étude effectuée en anglais, les nourrissons étaient familiarisés à des mots de type CVC (e.g., /kʌp/, « cup », tasse). Dans une seconde phase, différents passages étaient présentés ; ces dernières pouvaient soit contenir le mot familier (e.g., « cup ») soit contenir un pseudo-mot ne différant que par un seul phonème (e.g., /tʌp/). Leurs résultats montrent que les nourrissons écoutaient plus longtemps les passages contenant les mots familiers. Le signal acoustique de parole étant continu par nature et ne disposant donc pas de frontières évidentes (e.g., pauses) entre les mots (cf. Chapitre 2), ces résultats suggèrent qu'à l'âge de 7.5 mois, un nourrisson est capable de segmenter le signal acoustique de parole pour isoler les mots avec lesquels il a été familiarisé. A l'âge de 6 mois, les nourrissons sont également capables d'effectuer cette opération pour un mot inconnu, mais uniquement lorsque celui-ci est précédé par un mot familier (e.g., /mʌmi/, « Mommy », maman, Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Cela suggère qu'un signal fréquemment rencontré, dont un nourrisson possèderait une trace en mémoire, aide à segmenter le signal acoustique en unités discrètes (i.e., en mots). A l'appui de cette hypothèse, une étude effectuée par Kooijman et collègues (Kooijman, Hagoort, & Cutler, 2005) a mis en évidence que des nourrissons âgés de 10 mois présentaient une activité cérébrale spécifique (potentiels évoqués) à l'écoute d'un mot présenté dans le contexte d'une phrase, lorsque celui-ci avait préalablement été présenté isolément. Dans cette étude, la réalisation acoustique des items variait considérablement en fonction de leur mode de présentation, seul ou dans le contexte d'une phrase. Les résultats suggèrent donc que dès les premiers stades de son développement, l'enfant est capable de *mémoriser* les mots, c'est-à-dire *de s'en créer une représentation* en mémoire (voir e.g., Johnson, 2006 ; Goldinger, 1998 ; pour des hypothèses alternatives). Cette représentation serait suffisamment détaillée sur le plan phonologique pour lui permettre de reconnaître un mot malgré ses différentes réalisations acoustiques (voir e.g., Peperkamp & Dupoux, 2002) et d'en extraire son signal lorsque celui-ci est présenté dans le contexte d'une phrase (voir Houston, 2005; Werker, 2007, pour des revues). D'autres travaux (Newman, Ratner, Jusczyk, Jusczyk, & Dow, 2006) ont d'ailleurs montré que cette capacité à segmenter un mot dans une phrase semble être prédictive du développement de certaines capacités linguistiques évaluées à un âge postérieur. En effet, ces auteurs ont montré que cette

habileté, évaluée dans la première année de vie, permettait de prédire la taille du lexique expressif d'un grand nombre d'enfants à l'âge de 2 ans. En d'autres termes, cette capacité à segmenter le signal acoustique de parole en mots semble fortement liée au nombre de mots qu'un enfant est capable de produire l'année suivante (voir aussi Tsao, Liu, & Kuhl, 2004, pour des résultats similaires).

Egalement, plusieurs études ont cherché à évaluer la capacité des nourrissons à reconnaître le des mots familiers, présentés isolément (e.g., Hallé & de Boysson-bardies, 1994, 1996; Mandel, Jusczyk, & Pisoni, 1995; Swain, Zelazo, & Clifton, 1993). Swain et al. (1993) ont montré que lorsque des nouveau-nés âgés de 24h maximum sont familiarisés avec le signal acoustique d'un mot (e.g., /bɪːgl/, « beagle », beagle) ces derniers préfèrent écouter ce mot 24 heures plus tard, par rapport à un mot nouveau (e.g., /tɪndər/, « tinder », amadou). Ce résultat suggère donc que dès ses premiers jours de vie un enfant est capable de mémoriser un signal acoustique correspondant à un mot. A l'âge de 4.5 mois, il a également été montré que des nourrissons sont capables de différencier le signal acoustique correspondant à leur propre nom par rapport à d'autres prénoms (Mandel et al., 1995). Dans une étude effectuée en français, Hallé & de Boysson-Bardies (1994) ont cherché à savoir si des nourrissons âgés de 11 et 12 mois étaient capables de reconnaître des mots familiers (i.e., faisant partie des premiers mots prononcés par les enfants, e.g., «poupée ») et rares dans le langage oral de l'adulte (e.g., « diffus ») en l'absence de contexte linguistique adjacent (e.g., phrase). Les résultats montrent que les nourrissons préféraient écouter les mots familiers aux mots rares, quel que soit leur âge. Ces résultats suggèrent que les enfants sont capables de reconnaître un mot familier, en l'absence de contexte linguistique. Deux ans plus tard, ces mêmes auteurs (Hallé & de Boysson-Bardies, 1996), ont également présenté des mots familiers et rares à des nourrissons âgés de 11 mois, en l'absence de contexte linguistique adjacent. Dans cette étude les mots familiers étaient équivalents aux mots rares en termes d'indices prosodiques (i.e., correspondant à l'intonation, l'accent tonique, la durée de prononciation de certains phonèmes), ces indices acoustiques pouvant être utilisés pour la reconnaissance de mots chez le nourrisson (voir Peperkamp & Dupoux, 2002; Werker, 2007, pour des revues à ce sujet). Ces auteurs ont montré qu'à l'instar de leurs résultats précédents (Hallé & De Boysson-Bardies, 1994), les nourrissons de 11 mois prêtaient plus attention aux mots familiers qu'aux mots rares. Ceci restait vrai, même lorsque le VOT du premier phonème du mot familier (e.g., « poupée ») était modifié (e.g., « boupée »). Cette information suggère qu'à l'âge de 11 mois, même si un enfant dispose de représentations de mots dans son lexique, celles-ci ne sont pas encore extrêmement détaillées sur le plan phonologique.

Globalement, l'ensemble de ces résultats suggère que les représentations de mots en mémoire commenceraient à se construire très tôt dans le développement, dès la première année de vie (cf. Hallé & de Boysson Bardies, 1994, 1996 ; Dupoux & Peperkamp, 2002). Cependant, le fait que des nourrissons soient capables de reconnaître un pattern acoustique comme familier ne signifie pas forcément qu'ils peuvent y associer un sens. Certaines études montrent tout de même que des enfants, dès l'âge de 12-13 mois, seraient capables de comprendre la signification d'un nombre croissant de mots (Werker, 2007). Pour étudier cette hypothèse, Swingley & Aslin (Swingley & Aslin, 2000) ont présenté à des nourrissons de langue maternelle anglaise âgés de 18 à 23 mois deux images représentant des objets connus (e.g., image d'un bébé vs. image d'un chien). Le signal acoustique correspondant au nom de l'une ou l'autre des images était présenté simultanément (e.g., /beɪbi/, « baby »). Ce mot était inséré dans une phrase porteuse de type « Look at [mot] » (« Regarde le [mot] »). Leurs résultats indiquent que les bébés dès l'âge de 18 mois regardaient plus rapidement et plus longtemps l'image correspondant à ce qui était entendu. Cela suggère que les enfants sont bien capables d'associer un sens à un signal acoustique, au moins dès leur deuxième année de vie. Par ailleurs les nourrissons regardaient plus rapidement l'image correspondant au mot entendu lorsque celui-ci était correctement prononcé (e.g., « baby ») plutôt que lorsque le premier phonème était modifié sur le plan phonétique (e.g., /veɪbi/, « vaby »). Cette information suggère que dès l'âge de 18 mois, un enfant dispose de représentations relativement détaillées sur le plan phonologique.

En résumé, l'ensemble de ces résultats suggère que les enfants construisent des représentations lexicales et que ces représentations de mots sont associées à un sens, avant l'âge de deux ans. L'objectif de la prochaine section consiste à passer en revue différentes études s'intéressant à l'influence de ces représentations sur le processus de reconnaissance de mots, chez l'enfant.

5.1.2.2. Influence de l'information lexicale chez l'enfant

Alors que de nombreux travaux se sont intéressés au processus de reconnaissance de mots chez le nourrisson (cf. Werker, 2007, pour une revue), un plus petit nombre de recherches semble avoir examiné le processus de reconnaissance de mot parlés chez l'enfant d'âge scolaire (voir Walley, 2005, pour une revue). Dans cette section, nous allons nous intéresser aux travaux qui ont étudié l'influence de l'information lexicale sur le décodage du signal de parole en modalité auditive (voir section 2.2.5 du Chapitre 2, pour des références

chez l'adulte) chez l'enfant d'âge scolaire (i.e., de 5 à 11 ans : Ackroff, 1981; Garlock, Walley, & Metsala, 2001; Krull, Choi, Kirk, Prusick, & French, 2010; Mani & Plunkett, 2011; Metsala, 1997; Walley, 1988, 1993; Wang, Wu, & Kirk, 2010). Citons premièrement les travaux de Walley (1988) qui ont été parmi les premiers à étudier cette question. Dans cette étude, le paradigme de restauration phonémique a été utilisé (Warren, 1970, cf. Chapitre 2). Rappelons que ce paradigme consiste à rajouter (condition « rajoutée ») ou à remplacer (condition « remplacée ») une portion d'un mot correspondant à un phonème par du bruit. Lorsque l'on demande à des participants de juger si, à travers ce bruit, le phonème est présent ou absent, ces derniers ont tendance, en condition « remplacée », à percevoir le phonème, alors même que celui-ci est en réalité absent du signal acoustique. On dit qu'ils « restaurent » automatiquement le segment manquant. Or Samuel (1981) a montré que chez des adultes, cet effet de restauration phonémique avait principalement lieu dans des mots (e.g., "pr**o**gress", /pr**əʊ**grɛs/[43]) plutôt que dans des pseudo-mots (e.g., "cr**o**gless", /kr**əʊ**glɛs/). Cela suggère que cet effet peut être dû, au moins en partie, à une influence de l'information lexicale. Dans son étude, Walley (1988) a étudié l'effet de restauration phonémique dans des mots, chez des adultes et des enfants âgés de 5 ans. Le phonème sur lequel portait l'illusion pouvait être soit situé en début, en milieu ou en fin de mot. Dans l'hypothèse que le début de mot joue un rôle proéminent dans le processus d'accès au lexique (Warren & Marslen-Wilson, 1987; 1988, cf. Chapitre 2), une détérioration du signal en début de mot devrait davantage gêner sa reconnaissance par rapport à une même perturbation située en position médiale ou finale. Cela devrait avoir pour conséquence une plus faible influence de l'information lexicale et donc une diminution de l'effet de restauration phonémique dans la première condition (position initiale) par rapport aux autres (position médiale et finale). Les résultats montrent que globalement, les enfants présentaient un plus faible effet de restauration phonémique que les adultes (cf. Ackroff, 1981, cité par Walley, 1988, pour des résultats similaires avec des enfants âgés de 6 et 8 ans). De plus, aucun effet de la position du phonème sur lequel portait l'illusion n'a été observé chez les enfants. Néanmoins, chez les adultes, conformément aux prédictions formulées par l'auteur, l'effet de restauration phonémique était plus faible lorsque celui-ci portait sur un phonème initial. Ces données suggèrent donc que l'effet de restauration phonémique observé chez les enfants de 5 ans n'était pas, à la différence de celui observé chez les adultes, influencé par l'information lexicale. Globalement, ces résultats suggèrent que l'information lexicale aurait

[43] La lettre en gras indique le phonème manquant

moins d'influence sur le processus de décodage phonémique/phonétique chez l'enfant que chez l'adulte.

Notons néanmoins que certaines études suggèrent que des caractéristiques inhérentes au niveau lexical (i.e., telles que la fréquence lexicale ou la densité du voisinage phonologique) influenceraient le processus de reconnaissance de *mots* (Garlock, et al., 2001; Krull, et al., 2010; Mani & Plunkett, 2011; Metsala, 1997; Wang, et al., 2010). En utilisant une tâche de *gating* (cf. Chapitres 1 et 2), Metsala (1997) a étudié la capacité de dénomination de différents mots, chez des participants adultes mais également chez des enfants âgés de 7, 9 et 11 ans. La fréquence lexicale des mots (fréquents vs. rares) et la densité de voisinage phonologique (dense vs. faible) ont été manipulées. Les performances ont montré que malgré le fait que les enfants avaient besoin de plus de signal que les adultes pour parvenir à identifier le mot présenté, les performances de l'ensemble des participants étaient sensibles à ces deux paramètres. Plus précisément, ces auteurs ont mis en évidence que les performances pour deviner un mot étaient meilleures lorsque le mot en question était fréquent ou lorsqu'il disposait d'une densité de voisinage faible. Ce résultat est en accord avec ceux trouvés chez l'adulte (cf. Luce & Pisoni, 1998, pour une revue). Garlock et al. (2001) ont pour leur part testé l'influence de ces deux paramètres avec une tâche de répétition de mots, chez des adultes et des enfants âgés de 5.5 et 7.5 ans. Ces derniers ont mis en évidence que les pourcentages de répétitions correctes étaient meilleurs pour les mots disposant d'une faible densité de voisinage phonologique. Néanmoins, aucun effet de la fréquence lexicale n'a été observé sur ces performances. Récemment, Krull et al. (2010) ont montré que des mots fréquents et disposant d'une densité de voisinage faible présentés avec différents niveaux de bruit (i.e., RSB = -2, 0, 2 et 4 dB) étaient également plus facilement identifiés par des enfants âgés de 5 à 12 ans (voir aussi Wang et al., 2010 ; pour des résultats similaires). Remarquons qu'une très récente étude (Mani & Plunkett, 2011) a même montré que les performances d'enfants âgés de 24 mois dans une tâche d'amorçage phonologique étaient également modulées par la densité du voisinage phonologique des stimuli.

En résumé, l'apport de l'ensemble de ces travaux peut se scinder en deux parties. Premièrement, nous avons vu que bien que l'information visuelle joue un rôle important dans la perception de la parole chez le nourrisson (e.g., Kuhl & Meltzoff, 1982 ; Patterson & Werker, 2003 ; Weikum et al., 2007) son influence reste relativement faible à des stades plus avancés dans l'enfance (e.g., Sekiyama & Burnham, 2008). Ensuite, alors même que des nourrissons semblent disposer de représentations de mots en mémoire dès leur première année de vie (e.g., Hallé & de Boysson-Bardies, 1996) ces représentations ne semblent pas

influencer la perception de l'information phonétique chez les enfants d'âge scolaire comme chez les adultes (Walley, 1988).

5.1.3. Objectifs et méthodes de l'Etude 5

L'objectif de la présente étude consiste à examiner l'influence respective du niveau lexical et de l'information visuelle sur le processus de reconnaissance de mots, chez des enfants d'âge scolaire (i.e., ayant de 5 à 10 ans). Les Etudes 1 et 2 présentées dans le Chapitre 3 ont montré à l'aide de tâches de détections de phonèmes que chez l'adulte, l'information visuelle, en présence d'une information auditive congruente et détériorée participe au processus d'activation des représentations lexicales. Pour cela, nous avons utilisé une tâche de détection de phonèmes consonantiques (Etude 1) et vocaliques (Etude 2) dans des mots et des pseudo-mots, présentés en modalité A et AV, à différents niveaux de détérioration du signal acoustique (sans bruit, -9 dB, -18 dB). Les résultats montrent qu'en présence de bruit dans le signal acoustique (i.e., à -9 dB et à -18 dB), cette différence était plus importante en modalité audiovisuelle qu'en modalité auditive. Ce résultat indique que l'information visuelle participe à l'activation des unités lexicales et ce majoritairement lorsque l'information auditive est détériorée.

L'objectif de cette étude est d'observer si les enfants, comme les adultes, sont capables d'exploiter visuellement les mouvements articulatoires de leur interlocuteur dans le processus d'activation des représentations lexicales. Pour cela, nous avons décidé d'utiliser une tâche de détection de phonèmes vocaliques. Nous avons choisi d'employer des phonèmes-cibles vocaliques plutôt que consonantiques pour plusieurs raisons. Premièrement, il semblerait que les voyelles jouent, par rapport aux consonnes, un rôle important dans le développement de la parole (Locke, 1993). Ensuite, nous voulions rendre la tâche la plus facile et la plus adaptée à l'étude des compétences des enfants, tout en conservant la possibilité de comparer ces performances avec celles observées chez l'adulte (Chapitre 3). Ainsi, nous avons sélectionné des cibles vocaliques plutôt que consonantiques car ces dernières sont plus saillantes sur le plan perceptif (Ladefoged, 2001) et parce que leur reconnaissance semble mieux résister à l'ajout de bruit dans le signal acoustique que leurs homologues consonantiques (Nooteboom & Doodeman, 1984, cité par Cutler et al. 2000). Nous avons donc sélectionné une tâche de détection de phonèmes vocaliques insérés dans des mots et des pseudo-mots. Ces derniers sont présentés en modalité A ou AV, avec détérioration du signal acoustique (i.e., à -9 dB) à des enfants âgés de 5 à 10 ans. De la même

manière que dans les Etudes 1 et 2 effectuées chez l'adulte, nous avons donc introduit du bruit blanc masquant dans le signal acoustique des stimuli. Cette manipulation a été effectuée afin (1) d'éviter un plafonnement des performances pour l'ensemble des conditions et (2) de maximiser l'apport de l'information visuelle en présence de l'information auditive. Une situation de perception de la parole dans le bruit a donc préférentiellement été sélectionnée plutôt qu'une situation ou les informations auditive et visuelle ne sont pas congruentes (McGurk & MacDonald, 1976). En effet, bien que nous ayons utilisé une détérioration artificielle du signal acoustique (i.e., bruit blanc), percevoir la parole dans un environnement bruité, comme dans une classe d'une vingtaine d'élèves, constitue une activité journalière pour l'enfant, alors qu'il existe très peu de situations quotidiennes où les informations auditives et visuelles ne sont pas congruentes (films doublés). Afin de pouvoir comparer les performances obtenues par les enfants à celles des adultes et de limiter la difficulté de la tâche, nous avons sélectionné le niveau de bruit le plus faible utilisé dans l'Etude 2 (i.e., -9 dB).

Nous faisons l'hypothèse que si le bénéfice de l'information visuelle augmente avec l'âge, nous devrions obtenir un avantage plus important dans la condition AV par rapport à la condition A pour les enfants les plus âgés par rapport aux plus jeunes (Sekiyama & Burnham, 2008). En d'autres termes, nous devrions obtenir uniquement un faible bénéfice lié à la présence de l'information visuelle pour les enfants âgés de 5 à 8 ans. Si l'influence de l'information lexicale sur la perception des phonèmes augmente avec l'âge nous devrions observer un effet de supériorité du mot plus important pour les enfants plus âgés. Enfin, si l'information visuelle contribue au processus d'accès au lexique dès l'enfance, nous nous attendons à observer un effet de supériorité du mot plus important en modalité AV qu'en condition A.

5.2. ETUDE 5 : ROLE DE L'INFORMATION VISUELLE ET LEXICALE DANS LE PROCESSUS DE RECONNAISSANCE DE MOTS CHEZ L'ENFANT

Fort, M., Spinelli, E., Savariaux, C. & Kandel, S. (en révision). Audiovisual word recognition in children. *International Journal of Behavioral Development.*

5.2.1. Méthode

5.2.1.1. Participants

Quatre-vingt-seize enfants âgés entre 5;2 et 10;10 ans ont participé à cette étude. Ils étaient repartis en cinq groupes différents selon des critères d'âge et d'appartenance scolaire : 5-6 ans, grande section de maternelle (âge moyen : 5;8, N = 19), 6-7 ans, cours préparatoire (âge moyen : 6;11, N = 18) ; 7-8 ans, cours élémentaire première année (âge moyen : 7;11, N = 20) ; 8-9 ans, cours élémentaire deuxième année (âge moyen : 8;11, N = 20) ; 9-10 ans, cours moyen première année (âge moyen : 9;10 mois, N = 19). Tous étaient de langue maternelle française. Aucun d'entre eux n'avait de déficit sur le plan auditif et tous avaient une vision normale ou corrigée. Aucun d'entre eux n'avait redoublé de classe dans leur cursus au moment de l'étude. Les participants ont été recrutés dans trois écoles de la région Iséroise, sous consentement parental mais également de l'Inspecteur d'Académie.

5.2.1.2. Stimuli

5.2.1.2.1. Items expérimentaux

Un corpus de 20 paires de mots/pseudo-mots a été sélectionné (cf. Annexe E) à partir du corpus utilisé pour l'Etude 2 (voir Chapitre 3, section 3.3.1.2). Aucune base ne répertoriant la fréquence des mots dans le langage oral chez l'enfant, nous nous sommes assurés, à l'aide d'une étude préalable effectuée sur 15 enfants âgés de 5 ans que l'ensemble des mots utilisés dans cette étude étaient reconnus dans 100 % des cas. Chacun de ces items comportait un des trois phonème-cibles vocaliques sélectionnés pour cette étude : la moitié de ceux-ci pouvaient engendrer soit un mouvement d'arrondissement (/o/, /y/) soit un mouvement d'étirement des lèvres (/e/).

5.2.1.2.2. Items de remplissage

Vingt paires de mot/pseudo-mot correspondant aux items de remplissage (i.e., ne contenant pas le phonème-cible à détecter) ont été sélectionnés à partir du corpus utilisé pour l'Etude 2 (cf. Chapitre 3). Ces items de remplissage étaient choisis afin que l'ensemble des items le composant diffère sur le plan acoustique et articulatoire avec le phonème-cible.

5.2.1.2.3. Enregistrement des stimuli

Voir section 3.3.1.2.3 du Chapitre 3.

5.2.1.3. Procédure

Les participants étaient évalués dans une pièce calme située à l'extérieur de leur salle de classe, dans l'enceinte de leur établissement. Ils étaient assis à 50 cm environ d'un écran d'ordinateur portable DELL (1024 x 768 pixels) et la composante vidéo des stimuli était présentée à une vitesse de 25 images par secondes. Le signal acoustique était joué à une fréquence d'échantillonnage de 44100 Hz à travers deux enceintes SONY SRS-88 situées de chaque côté de l'ordinateur. Chaque enfant était vu deux fois lors de deux sessions situées à une ou deux semaines d'intervalle. A l'intérieur de chaque session, la totalité des stimuli était présentée soit en modalité A (signal acoustique et visage fixe de la locutrice) soit en modalité AV (signal acoustique et visage de la locutrice en mouvement). Pour chaque session, les enfants devaient détecter le phonème-cible dans l'item présenté (mot ou pseudo-mot). Ils savaient que le phonème-cible pouvait être ou ne pas être dans l'item proposé. Il leur était demandé d'appuyer le plus rapidement possible sur la barre d'espace uniquement lorsque le phonème était présent (tâche Go/No Go). Avant chaque essai, l'expérimentateur s'assurait que la main de l'enfant était positionnée juste au-dessus de la barre d'espace. La main de réponse employée était la main dominante. L'ordre de la modalité de présentation était contrebalancé entre les participants Chaque item était donc présenté deux fois à un enfant : une fois en modalité auditive, une fois en modalité audiovisuelle. Afin de limiter la charge cognitive liée à la tâche, le type de voyelle à détecter (/o/, /y/ ou /e/) était présenté par bloc. En conséquence, le phonème-cible était présenté en modalité auditive une seule fois avant chaque bloc. A l'intérieur de chaque bloc, la moitié des items contenaient l'item à détecter (items expérimentaux) alors que l'autre moitié ne le contenait pas (items de remplissage). Il était demandé aux participants de tout aussi bien prêter attention au signal acoustique qu'à la composante vidéo des stimuli (e.g., Alsius et al., 2005). Une session d'entraînement de 6 items précédait chaque bloc et pouvait être répétée aussi souvent que

nécessaire. La génération des stimuli et la collecte du type et du temps de réponse était assurée par le logiciel E-Prime 2.0 (*Psychological Software Tools*, *Pittsburgh*, *PA*). La durée de passation était de 15 minutes pour chaque session.

5.2.2. Résultats

Le pourcentage de détections correctes ainsi que la moyenne des temps de réponse (mesurés à partir du début du phonème-cible) sur les détections correctes ont été calculés pour chaque participant et chaque paire d'items. Une analyse de la variance (ANOVA) 5 (Age : 5-6 ans vs. 6-7 ans vs. 7-8 ans vs. 8-9 ans vs. 9-10 ans) x 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) a donc été effectuée en échantillon appariés, sur les temps de réponse d'une part et les détections correctes, par participants (F_1) et par items (F_2). Les temps de réponse considérés comme aberrants (i.e., inférieur à 100 ms ou supérieur à 2500 ms) ou se situant à plus de 2 écart-types de la moyenne du participant pour chaque condition respective ont été exclus de l'analyse. Suite à cette opération 5 % des données totales a été écarté.

5.2.2.1. Détections correctes

Les pourcentages de détections correctes pour les conditions bruitées de l'Etude 5 sont présentés dans la Figure 31.

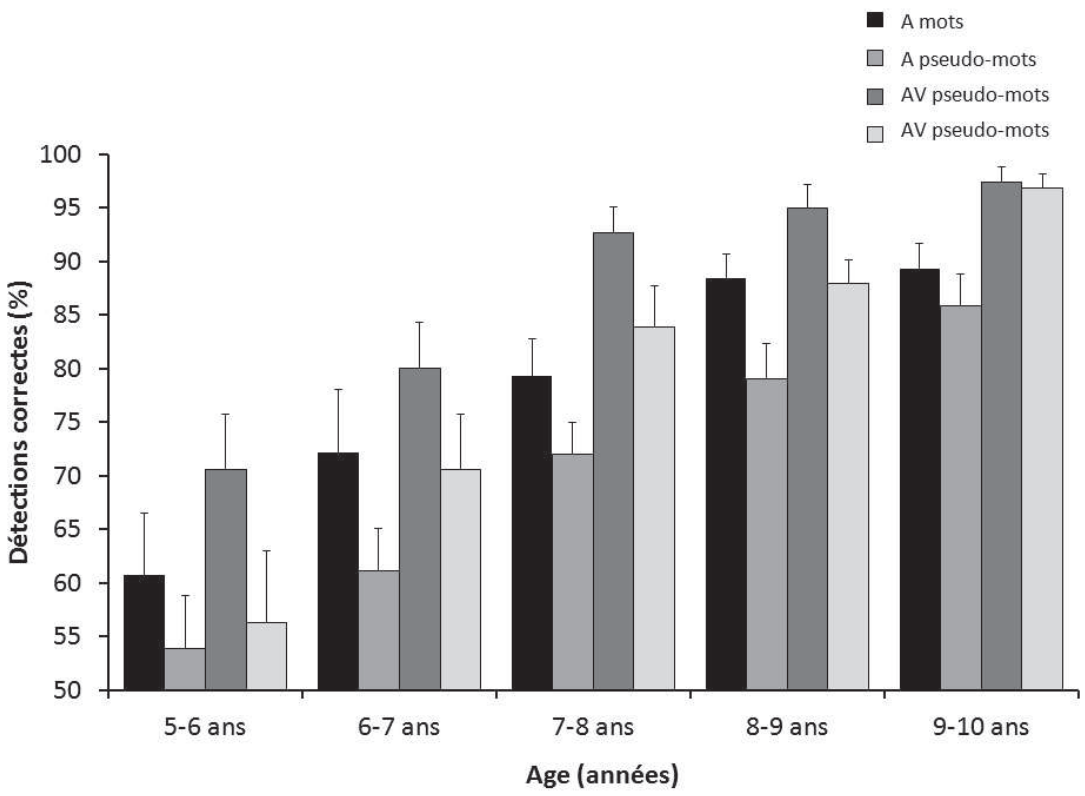


Figure 31. *Pourcentage de détections correctes à-9 dB pour les conditions auditive seule (A) et audiovisuelle (AV) de l'Etude 5. Les barres d'erreurs représentent l'erreur type.*

L'analyse statistique des données a permis de mettre en évidence un effet principal de l'âge, $F_1(4, 91) = 18.58$, $p < .001$, $\eta^2p = .45$, $F_2(4, 19) = 43.16$, $p < .001$, $\eta^2p = .65$, les résultats augmentant linéairement avec l'âge, $F_1(4, 91) = 71.61$, $p < .001$, $\eta^2p = .44$, $F_2(4, 19) = 155$, $p < .001$, $\eta^2p = .62$. Un effet principal de la modalité a également été obtenu, $F_1(4, 91) = 31.59$, $p < .001$, $\eta^2p = .26$, $F_2(4, 19) = 52$, $p < .001$, $\eta^2p = .35$, révélant globalement de meilleures performances pour la condition AV que pour la condition A. L'effet principal du Statut Lexical était également significatif, $F_1(4, 91) = 53.20$, $p = .001$, $\eta^2p = .37$, $F_2(1, 19) = 33.91$, $p < .001$, $\eta^2p = .26$, signifiant que les participants ont un pourcentage de détections correctes plus élevé lorsque le phonème-cible dans un mot plutôt que dans un pseudo-mot (i.e., effet de supériorité du mot). Aucun effet d'interaction simple ou double n'a été observé entre ces facteurs.

5.2.2.2. d'

Afin de prendre en compte le pourcentage de fausses alarmes dans nos analyses, un *d'* a été calculé pour chaque participants. Une analyse de la variance (ANOVA) 5 (Age : 5-6 ans vs. 6-7 ans vs. 7-8 ans vs. 8-9 ans vs. 9-10 ans) x 2 (Modalité de Présentation : A vs. AV) x 2 (Statut Lexical : mot vs. pseudo-mot) a donc été effectuée par participants (F_1) sur cet indice.

De la même manière que sur les détections correctes, l'analyse a permis de mettre en évidence un effet de l'âge, $F_1(4, 91) = 14.88$, $p < .001$, $\eta^2p = .40$. De meilleures performances ont été observées en modalité AV par rapport à la modalité A, $F_1(4, 91) = 160.22$, $p < .001$, $\eta^2p = .64$. Un effet de supériorité du mot a également été obtenu sur les *d'*, F_1 (4, 91) = 19.63, $p < .001$, $\eta^2p = .18$. Par ailleurs, aucun effet d'interaction simple ou double n'a été observé entre ces facteurs.

Notons que les résultats obtenus sur les d' étant identiques à ceux obtenus sur les réponses correctes, ils ne seront pas spécifiquement commentés dans la suite de cette étude.

5.2.2.3. Temps de réponse

Les temps de réponse moyens pour la condition bruitée de l'Etude 5 sont présentés dans la Figure 32.

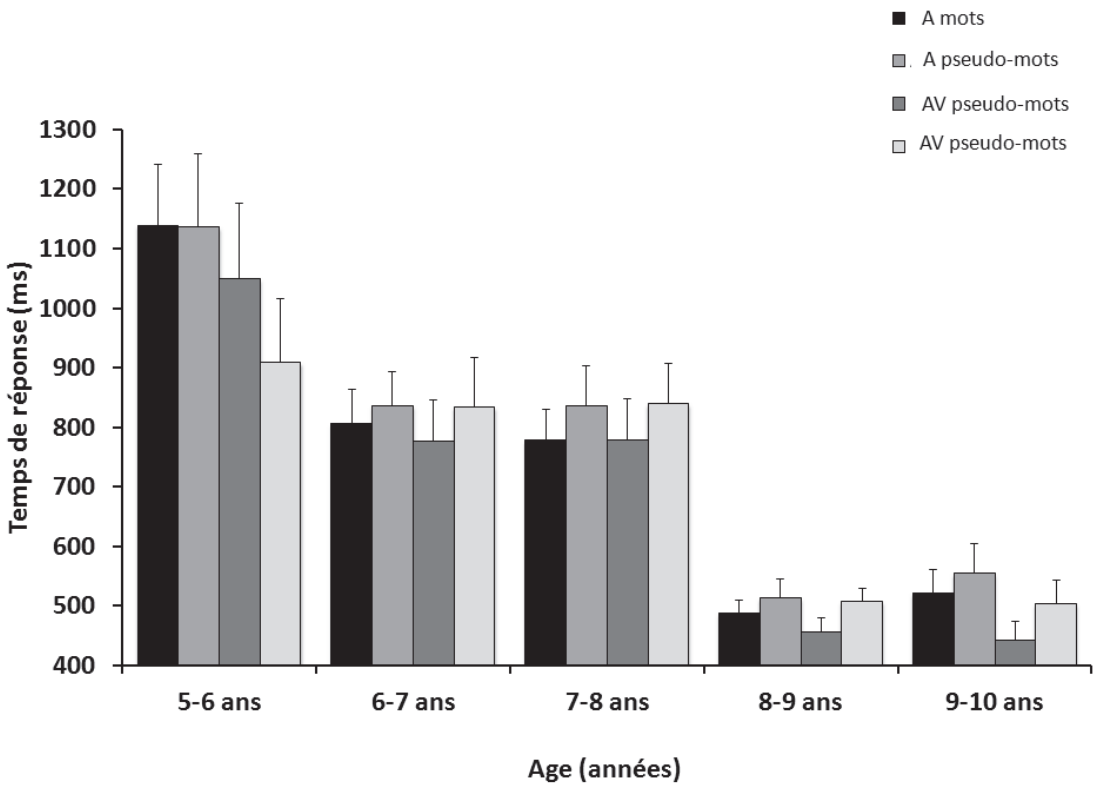


Figure 32. *Temps de réponse moyens (en millisecondes, ms) à -9 dB pour la condition auditive (A) et audiovisuelle (AV) de l'Etude 5. Les barres d'erreurs représentent l'erreur type.*

L'analyse statistique a mis en évidence un effet principal de l'âge, $F_1(4, 91) = 15.01$, $p < .001$, $\eta^2p = .40$, $F_2(4, 19) = 81.57$, $p < .001$, $\eta^2p = .77$. L'effet principal du statut lexical était seulement significatif par items, $F_1(4, 91) = 1.75$, $p > .05$, $\eta^2p = .02$, $F_2(1, 19) = 6.4$, $p = .01$, $\eta^2p = .06$. Bien qu'aucune interaction n'ait été obtenue entre le statut lexical et l'âge, $F_1(4, 91) = 1.33$, $p > .05$, $\eta^2p = .40$, $F_2 < 1$, des comparaisons par paires ont été effectuées afin de tester l'effet de supériorité du mot pour chaque groupe d'âge dans l'analyse par items. Les résultats ont révélé que l'effet de supériorité du mot n'était significatif que pour les 9-10 ans, $F_2(1, 19) = 4.94$, $p < .05$, $\eta^2p = .06$, mais pas pour les autres groupes d'âge ($0 < F_2 < 1.6$, $p > .05$). Aucun effet principal de la modalité n'a été obtenu, $F_1(4, 91) = 3.9$, $p > .05$, $\eta^2p = .04$, $F_2 < 1$. Aucune autre différence ne s'est révélée significative (tous les $F_1 < 1$). Par ailleurs, aucun autre effet d'interaction simple ou double n'a été observé entre ces facteurs.

5.2.3. Discussion

L'objectif de l'Etude 5 était d'examiner l'apport de l'information visuelle et lexicale au processus de reconnaissance de mots dans une perspective développementale. Dans ce travail, nous avons exploré le rôle de ces deux types d'informations chez des enfants âgés de 5 à 10 ans. Pour cela, à l'instar de l'Etude 2 effectuée chez l'adulte, nous leur avons proposé une tâche de détection de phonèmes vocaliques, présentés dans des mots ou des pseudo-mots, en modalité A et AV, avec du bruit dans le signal acoustique (-9 dB).

Les données montrent tout d'abord un effet de l'âge sur les performances. Ensuite, les participants ont obtenus des scores plus élevés en modalité AV qu'en condition A. Egalement, de meilleures performances ont été observées pour détecter un phonème dans un mot plutôt que dans un pseudo-mot. Aucun effet d'interaction n'a été observé entre ces différents facteurs.

5.2.3.1. Influence de l'âge

Les résultats de l'Etude 5 mettent en évidence une influence générale de l'âge des participants sur les performances. Les données sur les pourcentages de réponses correctes indiquent que celles-ci augmentaient linéairement avec l'âge. Or, les résultats ne mettent en avant aucune influence différentielle de l'âge en fonction de la présence ou non d'information visuelle ou la présence ou non d'information lexicale. Cet effet principal ne peut donc être justifié par le fait qu'un individu améliore le traitement de ces deux types d'informations au cours de son développement. Il pourrait premièrement être expliqué par une amélioration de la conscience phonémique des enfants avec l'âge, cette compétence leur étant nécessaire pour

effectuer la tâche (i.e., tâche de détection de phonèmes). La maturation chez l'enfant des structures corticales et sous-corticales impliquées dans le filtrage du signal sonore dans un environnement bruité (e.g., Anderson, Skoe, Chandrasekaran, Zecker, & Kraus, 2010) pourrait aussi rendre compte d'une partie de l'augmentation globale des performances avec l'âge. D'autres travaux devront être effectués afin de tester ces hypothèses.

5.2.3.2. Apport de l'information visuelle

Les résultats indiquent clairement que les participants obtiennent de meilleurs scores en modalité AV qu'en condition A, dès l'âge de 6-7 ans. A l'âge de 5-6 ans, un bénéfice lié à la présence de l'information visuelle a été observé uniquement pour les mots (M_{AV-A} = 9.8 %) pour les réponses correctes. De plus, les résultats observés sur l'indice *d'* montrent également un apport de l'information visuelle pour les 5-6 ans, pour les mots comme pour les pseudo-mots. Enfin, contrairement à d'autres études (e.g., Sekiyama & Burnham, 2008) aucune interaction entre la modalité et l'âge n'a été obtenue, laissant supposer que le bénéficie lié à la présence de l'information visuelle n'augmente pas significativement avec l'âge.

En conséquence, ces résultats suggèrent fortement que dès l'âge de 5-6 ans, les enfants sont capables d'utiliser l'information visuelle afin d'augmenter l'intelligibilité du signal acoustique, lorsque l'information auditive est détériorée. A l'instar de l'explication avancée pour les adultes, ce bénéfice peut être expliqué par le fait que lorsque l'information auditive est détériorée, le signal visuel devient complémentaire de l'information auditive en termes de type d'information véhiculée (Summerfield, 1987, cf. Chapitre 1). Parce que cet avantage était de taille similaire à travers les différents âges testés, nos résultats indiquent que les enfants possèdent, dès l'âge de 5-6 ans, une bonne capacité à décoder et utiliser le geste articulatoire afin d'augmenter l'intelligibilité des phonèmes, en présence d'une information auditive congruente et détériorée. Cette étude est, à notre connaissance, la première à mettre en évidence un apport aussi important de la modalité visuelle, à l'âge de 5-6 ans. En effet, Sekiyama et Burnham (2008) ont montré à l'aide d'une tâche de détection de syllabes que les enfants anglais australiens et japonais âgés de 6 ans ne bénéficiaient que faiblement de l'influence de cette modalité, en situation bruitée à -4 dB. Or les résultats de l'Etude 5 indiquent avec une tâche de détection de phonèmes que des enfants francophones, dès l'âge de 5-6 ans, bénéficient fortement de la présence de l'information visuelle, lorsque le signal acoustique est détérioré (à -9 dB).

5.2.3.3. Apport de l'information lexicale

Les résultats de l'Etude 5 mettent également en évidence un effet robuste du statut lexical pour l'ensemble des âges testés, à la fois pour les pourcentages de réponses correctes, les *d'*. Ce résultat indique que les participants avaient de meilleurs scores pour détecter des phonèmes-cibles vocaliques situés dans des mots plutôt que dans des pseudo-mots (i.e., effet de supériorité du mot). Cette étude est, à notre connaissance, une des premières à mettre en évidence cet effet chez l'enfant en reconnaissance de mots parlés (voir e.g., Grainger, Bouttevin, Truc, Bastien, & Ziegler, 2003, en reconnaissance de mots écrits). Les modèles d'accès au lexique élaborés chez l'adulte tels que Merge ou TRACE expliquent ce phénomène en postulant une diffusion d'activation du niveau lexical vers le niveau de décision phonémique. Ces données suggèrent donc que l'information lexicale influence le processus de décision phonémique également chez l'enfant au moins dès l'âge de 5-6 ans. Cela implique que les représentations phonologiques des mots connus par les enfants sont suffisamment détaillées dès l'âge de 5-6 ans pour influencer leur perception. Etant donné que ces performances ont été observées alors que le signal acoustique était détérioré, cela suggère que les enfants sont également capables de se servir du contexte lexical afin d'augmenter l'intelligibilité des phonèmes lorsque le signal de parole est ambigu.

Aucun effet d'interaction n'a été observé entre l'effet lexical et l'âge des participants. Or, le nombre de mots connus par un individu est très fortement corrélé à son avancée en âge, au moins durant l'enfance (e.g., pour les enfants âgés entre 3 et 8 ans, Edward et al., 2004). Si l'on suppose que les représentations phonologiques des mots connus par des enfants se spécifient sous l'influence de l'augmentation de la taille de leur lexique (Fowler, 1991, cité par Walley, 1993), (voir aussi, Garlock, 2001 ; Metsala, 1997 ; Walley, 1993, pour des propos similaires), ces dernières deviendraient plus détaillées avec l'âge. Dans cette hypothèse, nous aurions pu observer une influence du niveau lexical sur le niveau de traitement des phonèmes plus importante avec l'avancée en âge (et donc avec l'augmentation de la taille du lexique). En effet, bien que le pic d'expansion du vocabulaire d'un enfant arrive plus tôt dans son développement (i.e., autour de 18-24 mois, e.g., Dapretto & Bjork, 2000), il a été évalué que le nombre de mots entre 5 et 6 ans se situe entre 2500 et 5000 mots et augmente en moyenne de 3000 mots par an (Beck & McKeown, 1991). Cette augmentation reste non négligeable et devrait, toujours selon cette hypothèse, influencer la perception des phonèmes.

Nos résultats ne concordent pourtant pas avec ces prédictions. Cela pourrait s'expliquer tout d'abord par le fait que les représentations phonologiques des enfants sont suffisamment détaillées à cet âge (voir aussi e.g., Best, Tyler, Gooding, Orlando, & Quann,

2009; Saffran, Aslin, & Newport, 1996, pour des hypothèses alternatives). Ensuite, cette absence d'interaction peut également s'expliquer par le choix des phonèmes-cibles que nous avons utilisé. Nous avons employé une tâche de détection de phonèmes-cibles vocaliques afin de faciliter au maximum la tâche donnée aux enfants. En effet, les voyelles sont, comparativement aux consonnes, relativement saillantes dans le signal de parole (Ladefoged, 2001) et leur reconnaissance semble mieux résister à l'ajout de bruit dans le signal acoustique que leurs homologues consonantiques (Nooteboom & Doodeman, 1984, cité par Cutler et al. 2000). Il serait donc possible que même les enfants du plus jeune groupe d'âge (5-6 ans) disposent de représentations phonologiques (déjà) très détaillées pour les voyelles et qu'ainsi, aucun impact de l'expansion du lexique ne soit observé sur cette mesure.

5.2.3.4. Apport combiné de l'information visuelle et lexicale

Bien que ce travail indique que les enfants peuvent traiter et de bénéficier séparément de l'apport des informations visuelle et lexicale, nos données suggèrent qu'ils ne sont pas encore capables d'exploiter conjointement ces deux types d'indices pour effectuer la tâche. Effectivement, contrairement à nos attentes, aucune interaction significative entre le statut lexical et la modalité de présentation des items n'a pu être obtenue. Cela indique que, contrairement aux données observées chez l'adulte (sur le pourcentage de réponses correctes dans l'Etude 1, les temps de réponses dans l'Etude 2) l'effet de supériorité du mot n'était pas plus important lorsque le visage en mouvement de la locutrice était visible (condition AV) par rapport à une situation où seul le signal acoustique était disponible (condition A). En conséquence, il semblerait que l'information visuelle peut seulement activer les unités pré-lexicales mais qu'elle ne participe pas encore à l'activation des représentations lexicales, jusque dans les dernières phases de l'enfance (i.e., 9-10 ans). Il semblerait donc que les enfants, dès l'âge de 5-6 ans, soient capables de traiter l'information visuelle et lexicale séparément mais qu'ils ne puissent pas les combiner (au moins jusque l'âge de 9-10 ans) afin d'optimiser le processus de reconnaissance de mots. Les limites et perspectives de cette étude seront discutées dans la section 6.1.4 Chapitre 6.

CHAPITRE 6. DISCUSSION GENERALE, PERSPECTIVES ET CONCLUSIONS

« Pour les étudiants, la thèse de doctorat est le couronnement des études ; pour les enseignants-chercheurs, elle est la fin d'une étape et le début d'une autre »

Michel Beaud. (Beaud, 2006) p. 10.

6.1. DISCUSSION GENERALE

6.1.1. Rappel des principaux résultats

L'objectif des études présentées dans le manuscrit était d'étudier le rôle de l'information visuelle dans le processus de reconnaissance de mots, autrement dit dans le processus d'activation des représentations lexicales. Nous avons premièrement examiné cette question chez l'adulte, à l'aide d'une tâche de détection de phonèmes, en présence d'une information auditive congruente (Etudes 1 et 2). Cette information auditive pouvait être intacte ou bruitée. Nous avons montré que la présence de l'information visuelle permettait d'accélérer le processus de détection de phonèmes en présence de bruit (Etude 2) comme en son absence dans le signal acoustique (Etudes 1 et 2), par rapport à une situation où seule l'information auditive est présente. Ensuite, nous avons également mis en évidence une influence du niveau lexical sur le processus de détection de phonèmes plus importante en modalité AV qu'en situation A et ce uniquement lorsque le signal acoustique est détérioré. Ces résultats suggèrent donc que l'information visuelle participe au processus d'activation des représentations lexicales et ce particulièrement lorsque la reconnaissance d'un mot est difficile (i.e., en situation bruitée). Dans un deuxième temps, nous avons étudié l'influence de l'information visuelle seule dans le processus de reconnaissance de mots en utilisant un paradigme d'amorçage phonologique partiel (Etude 3 et 4). Nous avons mis en évidence que la seule présentation du geste articulatoire pour une amorce monosyllabique (e.g., articulation silencieuse de « bu ») facilitait la reconnaissance ultérieure d'un mot-cible bisyllabique (e.g., « bureau ») présenté auditivement (Etude 3 et 4). De plus, nous avons montré que cet effet d'amorçage était modulé en fonction de la fréquence et de la densité de voisinage phonologique de la cible, suggérant que le locus de cette facilitation était lexical plutôt que pré-lexical (Etude 4). En d'autres termes, ce résultat indique que le fait de voir le geste articulatoire correspondant au début d'un mot constitue une information suffisante pour contacter les représentations lexicales. Or, nous avons également montré que les effets d'amorçage obtenus étaient significatifs uniquement pour les mots-cibles de basses fréquence et disposant d'une densité de voisinage phonologique élevée. Ce résultat suggère donc également que l'apport de l'information visuelle au processus de reconnaissance de mots chez l'adulte serait particulièrement remarquable lorsque l'accès au lexique est difficile. Dans la dernière partie de ce manuscrit, nous avons étudié l'apport de l'information visuelle au processus d'activation des représentations lexicales en modalité audiovisuelle, chez des enfants âgés entre 5 et 10 ans. De la même manière qu'avec les participants adultes, nous avons étudié cette question, à l'aide d'une tâche de détection de phonèmes en présence d'une

information auditive congruente et bruitée. Les résultats ont montré une influence du niveau lexical et de l'information visuelle sur le processus de détection de phonèmes, quel que soit le groupe d'âge. Ces données suggèrent donc que dès l'âge de 5-6 ans, les enfants ont la capacité d'utiliser l'information lexicale mais aussi visuelle pour détecter un phonème dans une situation bruitée. Néanmoins aucune interaction entre ces deux facteurs n'a été obtenue pour aucun des groupes d'âges étudiés, suggérant que ces derniers ne sont pas, contrairement aux adultes, capables de combiner ces deux types d'information pour effectuer la tâche. En d'autres termes, ces résultats laissent supposer que, chez l'enfant, jusque l'âge de 10 ans, l'information visuelle serait uniquement décodée à un niveau pré-lexical.

En conclusion, comme nous l'avons fait remarquer dans le Chapitre 2, les modèles psycholinguistiques actuels décrivent le processus de reconnaissance de mots uniquement en modalité auditive. Autrement dit, seul le signal acoustique du signal de parole est considéré comme une entrée sensorielle possible pour accéder aux représentations lexicales. Or, nous avons montré dans ce travail que l'information visuelle joue également un rôle[44] dans ce processus, en présence d'une information auditive congruente (Etudes 1 et 2) ainsi qu'en l'absence de tout signal acoustique (Etude 3 et 4) chez l'adulte. L'objectif de la partie suivante consiste à étudier comment les modèles décrivant l'accès au lexique pourraient rendre compte de nos résultats si ces derniers incluaient l'information visuelle dans leur architecture.

6.1.2. Conséquences des résultats des Etude 1 et 2 pour les modèles d'accès au lexique

Dans les deux prochaines sections, nous allons décrire comment les modèles TRACE et Merge pourraient rendre compte des résultats obtenus avec les tâches de détection de phonèmes (Etude 1 et 2), si ces derniers considéraient le signal visuel de parole comme une entrée sensorielle. Remarquons que nous avons choisi de nous restreindre à ces deux modèles car ils prévoient des mécanismes spécifiques et opposés l'un à l'autre pour rendre compte de l'influence du niveau lexical sur le processus de détection et d'identification de phonèmes. Remarquons également que comme Merge est en réalité une version de Shortlist

[44] Dans une perspective simplificatrice, notons que nous avons considéré que l'influence de l'information visuelle serait équivalente pour l'ensemble des individus. Cependant, de nombreuses recherches indiquent que le traitement de l'information visuelle seule ainsi que son intégration à l'information auditive était sujette à une grande variabilité interindividuelle (genre : Strelnikov et al., 2008 ; présence d'un déficit auditif : Rouger et al., 2007 : âge et culture : e.g., Sekiyama & Burnham, 2008) pouvant largement complexifier les mécanismes décrits dans l'ensemble de ce manuscrit.

A spécifique aux tâches portant sur la détection ou d'identification de phonèmes, les conclusions que nous effectuerons pour Merge seront également valables pour Shortlist A.
Notons enfin qu'une des conséquences directe de l'incorporation du signal visuel de parole (en présence de l'information auditive) à ces modèles consiste à déterminer comment ces derniers pourraient rendre compte de l'intégration audiovisuelle de la parole dans leur architecture. Ce point ne faisant pas directement partie de nos hypothèses de recherches, nous en discuterons dans les perspectives (cf. section 6.2.3).

6.1.2.1. Interprétation des résultats des Etudes 1 et 2 pour le modèle TRACE

Selon le modèle TRACE (McClelland & Elman, 1986, voir section 2.3.2 du Chapitre 2 pour plus de détails à ce sujet), l'arrivée du signal de parole va premièrement générer de l'activation au niveau des traits. A ce stade, les différentes caractéristiques acoustico-articulatoires de base composant le signal sont extraites (place d'articulation, mode d'articulation, présence de voisement, etc.). Si le traitement de l'information visuelle devait être pris en compte dans son architecture, nous proposons qu'elle le soit dès le niveau des traits. En effet, nous avons vu dans le Chapitre 1 qu'un individu était capable d'extraire de l'information phonétique du signal visuel de parole (e.g., la place d'articulation, Smeele, 1994). Ainsi, la présence d'un signal visuel permettrait d'activer de manière plus importante ce niveau des traits, par rapport à une situation de perception de la parole en modalité auditive seule. Plus précisément, nous postulons que l'information visuelle devrait spécifiquement avantager les unités correspondant à des caractéristiques articulatoires visibles (comme la place d'articulation) plutôt que difficilement visibles (comme le voisement) (e.g., Fisher, 1968 ; Summerfield, 1987). De même, cette différence devrait être surtout marquée pour les unités codant des traits saillants dans le signal visuel de parole (e.g., comme une place d'articulation bilabiale pour un /p/, articulée à l'avant du conduit vocal), plutôt que pour des traits moins visibles (comme une place d'articulation vélaire pour un /k/, articulée plutôt à l'arrière du conduit vocal). Enfin, ces différences en termes d'activation devraient être spécialement marquées lorsque l'information auditive est détériorée, les traits articulatoires les plus saillants visuellement correspondant à ceux qui sont le plus facilement masqués par du bruit dans le signal acoustique (e.g., les traits pour la place d'articulation ; Summerfield, 1987). Notons ici que pour la perception du phonème /p/, la présence de l'information visuelle participera principalement à l'activation des traits « occlusive » et « bilabiale », la présence/absence de voisement étant une caractéristique difficilement visible dans le signal visuel de parole (Summerfield, 1987).

A la suite du niveau des traits, selon TRACE, l'activation va diffuser vers le niveau phonémique. De la même manière que pour le niveau des traits, la présence de l'information visuelle devrait permettre aux unités correspondant aux phonèmes hautement visibles sur le plan articulatoire (e.g., /p/) de recevoir plus d'activation que pour des phonèmes articulés un peu moins en avant, voire à l'arrière du conduit vocal (e.g., /t/, /g/), par rapport à une situation où seul le signal acoustique est disponible. Cette différence devrait également être plus marquée lorsque le signal acoustique est détérioré, permettant de rendre compte du fait que les phonèmes sont plus intelligibles en modalité AV qu'en condition A (e.g., Benoît et al., 1994 ; Binnie et al., 1974). Ces deux derniers postulats nous permettent d'expliquer les résultats de l'Etude 1, indiquant que les participants étaient plus rapides et plus performants pour détecter des phonèmes en modalité AV qu'en modalité A. En effet, il s'est avéré que cet effet était (1) principalement observé pour la détection de phonèmes consonantiques ayant une place d'articulation labiale (ou bilabiale) plutôt qu'alvéodentale et (2) uniquement lorsque le signal acoustique est bruité. Dans l'Etude 2, les résultats ont indiqué que les participants étaient plus rapides pour détecter un phonème vocalique en modalité AV qu'en modalité A, en présence comme en l'absence de bruit dans le signal acoustique. Dans le cadre du modèle TRACE, cela suggère également que la présence d'information visuelle permet d'augmenter l'activation reçue par le niveau des traits et celui des phonèmes, même lorsque le signal acoustique est intact.

Lorsque le stimulus est un mot, le flux d'activation va ensuite se propager du niveau phonémique vers le niveau lexical, permettant de sélectionner la représentation du mot stockée en mémoire correspondant au(x) signal(aux) d'entrée. Nous postulons que la présence d'informations visuelles permettrait, par association à des unités phonémiques activées précédemment, d'activer plus fortement les mots composés de phonèmes saillants sur le plan articulatoire (e.g., /p/ dans « cha**p**eau ») que les mots composés de phonèmes moins visibles (e.g., /t/ dans « man**t**eau »). Cette différence d'activation devrait être observée toutes caractéristiques lexicales étant équivalentes par ailleurs[45] (telles que la fréquence d'occurrence ou la densité du voisinage phonologique), ces paramètres étant également connus pour influencer le traitement de l'information visuelle à un niveau lexical (e.g., Mattys et al., 2002 ; voir Chapitre 3 et 4 pour plus de discussion à ce sujet).

Afin de rendre compte de l'effet de supériorité du mot, rappelons que TRACE postule l'existence d'un mécanisme rétroactif permettant de renvoyer de l'activation du niveau

[45] Notons ici que les mots « chapeau » et « manteau » disposent d'une densité de voisinage phonologique et d'une fréquence d'occurrence dans le langage oral relativement similaires (N = 18, F = 48.61 vs. N = 13, F = 36.16).

lexical vers le niveau des phonèmes. Dans ce modèle, l'activation peut diffuser de manière *bidirectionnelle* entre le stade phonémique et le stade lexical. Ce mécanisme de rétroaction a pour conséquence que les unités phonémiques -correspondant aux différents phonèmes du signal de parole- reçoivent plus d'activation lorsque ces dernières sont présentées dans le contexte d'un mot plutôt que d'un pseudo-mot, le niveau lexical n'existant pas pour le décodage des pseudo-mots. C'est grâce à ce supplément d'activation des unités phonémiques en contexte lexical que TRACE peut expliquer l'effet de supériorité du mot sur le plan comportemental.

En conséquence, si TRACE incorporait l'information visuelle (comme source d'information) dans son architecture, l'effet de supériorité du mot observé dans les Etudes 1 et 2 résulterait alors d'un processus descendant, postulant un retour d'activation en provenance d'un niveau d'analyse supra-ordonné (i.e., le niveau lexical) vers un stade inférieur de traitement (i.e., au niveau phonémique). En effet, afin d'expliquer nos résultats, TRACE devrait postuler que les unités lexicales reçoivent plus d'activation en condition AV qu'en condition A, autrement dit que l'information visuelle *participe* à l'activation des représentations lexicales. Cette différence d'activation permettrait un retour d'activation plus important du niveau lexical vers le niveau des phonèmes pour la condition AV, permettant de rendre compte des résultats des Etudes 1 et 2.

6.1.2.2. Interprétation des résultats des Etudes 1 et 2 pour le modèle Merge

Afin de rendre compte de l'effet de supériorité du mot, rappelons que le modèle Merge (Norris et al., 2000 ; voir Chapitre 2, section 2.3.4 pour une description plus complète du modèle) diffère principalement du modèle TRACE par rapport au sens de diffusion de l'activation entre le niveau de traitement des phonèmes et lexical. En effet, Merge suppose que cette propagation d'activation ne peut s'effectuer que de manière *unidirectionnelle* du niveau lexical vers le niveau de décision phonémique et réfute l'existence de tout mécanisme rétroactif des représentations lexicales vers d'autres stades. Pour rendre compte de l'effet de supériorité du mot obtenu avec la tâche de détection de phonèmes, ce modèle postule l'existence d'un stade de décision phonémique indépendant des deux autres niveaux (i.e., des niveaux lexical et pré-lexical). Ce stade est uniquement dédié aux opérations qui nécessitent une analyse portant spécifiquement sur les phonèmes et n'est pas connecté en permanence avec les autres niveaux. En effet, les connections entre ce stade de décision phonémique et les autres nœuds ne sont opérationnelles que lorsque la situation requiert d'effectuer une opération de décision phonémique (e.g., durant une tâche de détection de phonèmes).

Si Merge incorporait l'information visuelle dans sa structure, nous postulons de la même manière que pour TRACE que les unités des différents niveaux (pré-lexicaux, de décision phonémiques et lexicaux) devraient recevoir plus d'activation en modalité AV qu'en modalité A et ce spécifiquement lorsque l'information auditive est détériorée. Nous faisons également l'hypothèse que certaines unités devraient être plus avantagées (en termes de quantité d'activation reçue) lorsque celles-ci codent pour des traits phonétiques particulièrement saillants dans le signal visuel de parole. Ces deux postulats nous permettrait d'expliquer que les phonèmes sont mieux et plus rapidement reconnus en condition audiovisuelle qu'en condition auditive (Etude 1 et 2) principalement lorsque l'information auditive est détériorée et pour des phonèmes étant articulés à l'avant du conduit vocal (e.g., /p/, Etude 1). Afin d'expliquer que l'effet de supériorité du mot observé est plus important en modalité AV qu'en situation A (Etude 1 et 2), Merge devrait postuler, de la même manière que TRACE que l'information visuelle *participe* à l'activation des unités lexicales. A la différence de TRACE cependant, Merge expliquerait cet effet en supposant que la présence de l'information visuelle, permettrait de fournir plus d'activation *ascendante* du niveau pré-lexical vers le niveau lexical qu'en modalité A. Ainsi, la connexion entre le niveau lexical et le stade de décision phonémique rendue opérationnelle par la demande de la tâche permettrait aux unités de décision phonémique de recevoir une quantité d'activation provenant du niveau lexical plus importante en modalité AV qu'en modalité A. Ce mécanisme peut ainsi rendre compte du fait que dans l'Etude 1 et 2, l'effet de supériorité du mot est plus important en modalité AV qu'en modalité A.

En conclusion, c'est en (1) postulant que l'information visuelle *participe*, en présence de l'information auditive, au processus d'activation des représentations lexicales et (2) incluant la gestualité oro-faciale comme source d'information dans son architecture, que le modèle Merge peut expliquer le fait que nous observons un effet de supériorité du mot plus important en modalité audiovisuelle qu'auditive (Etudes 1 et 2).

6.1.3. Conséquences des résultats de l'Etude 3 et 4 pour les modèles d'accès au lexique

Dans cette partie, nous allons étudier comment les modèles Cohorte II, NAM, TRACE et Shortlist A pourraient rendre compte des résultats obtenus avec des paradigmes d'amorçage phonologique partiel (Etude 3 et 4), si ces derniers considéraient le signal visuel de parole comme source d'information dans leur architecture. Les résultats des Etudes 3 et 4 ayant mis en évidence que l'information visuelle *seule* permet d'activer les représentations lexicales, nous discuterons des conséquences liées aux propriétés du signal visuel de parole dans le processus d'activation des représentations lexicales.

6.1.3.1. Interprétation des résultats des Etudes 3 et 4 pour le modèle de la Cohorte II

Pour le modèle de la Cohorte II, l'effet d'amorçage phonologique partiel (e.g., Spinelli, 1999 ; Spinelli et al., 2001) est expliquée par le fait que l'amorce auditive /vɛʁ/, « ver » a permis de générer l'activation du candidat lexical /vɛʁvɛn/, « verveine » lors de la formation de la cohorte initiale. L'intervalle de 50 ms entre la présentation de l'amorce et de la cible étant relativement court, l'unité lexicale de « verveine » n'aurait pas encore eu le temps d'être éliminé de la cohorte. Autrement dit, c'est parce que « verveine » serait toujours activée à la fin de la présentation de l'amorce syllabique « ver » qu'un effet de facilitation est observé.

Si le modèle de la Cohorte II incluait l'information visuelle dans son architecture, la présentation de l'articulation silencieuse pour l'amorce /by/ permettrait donc de générer une cohorte initiale. Les candidats la composant devraient disposer de caractéristiques phonémiques en *début* de mot compatibles avec les informations articulatoires de cette amorce. Certaines informations n'étant pas ou peu disponibles dans le signal visuel de parole (i.e., la présence de voisement, la nasalisation, e.g., Summerfield, 1987), percevoir l'articulation silencieuse de /by/ permettrait de générer dans la cohorte l'ensemble des unités lexicales commençant par /by/ mais également par /py/ et /my/ (ainsi que par /bu/, /pu/ et /mu/, le /y/ étant un sosie labial du /u/). En d'autres termes, cela signifierait que l'information visuelle seule génèrerait des cohortes initiales de taille plus importante que l'information auditive. L'ISI entre l'amorce et la cible étant toujours de 50 ms, les candidats activés n'aurait pas le temps d'être éliminés avant la présentation de la cible, facilitant sa reconnaissance ultérieure. Ce mécanisme permet d'expliquer les effets de facilitation obtenus en condition visuelle seule de l'Etude 3, mais également le fait que l'effet d'amorçage obtenu

dans cette condition était de plus petite taille que lorsque l'amorce était présentée en modalité A ou AV.

Dans l'Etude 4, nous avons également observé que l'effet d'amorçage facilitateur était majoritairement présent pour les mots-cibles de basse fréquence lexicale. Or le modèle de la Cohorte II suppose qu'un mot de haute fréquence n'a besoin que de peu d'activation pour être reconnu. Si les effets d'amorçage en modalité visuelle seule n'ont été observés que pour les mots rares, cela peut être dû à un effet plafond des performances de reconnaissance des cibles auditives de haute fréquence.

Toujours dans l'Etude 4, nous avons observé que l'effet d'amorçage était corrélé positivement avec la densité du voisinage phonologique. Or le modèle de la Cohorte II précise que les candidats lexicaux pouvant être activés par la présentation d'une amorce telle que /by/ doivent absolument partager le même début avec celle-ci (e.g., /byʁo/, « bureau ») pour pouvoir faire partie de la cohorte initiale et ainsi faciliter le traitement ultérieur de la cible. Cependant, la notion de voisinage phonologique désigne des mots partageant les mêmes phonèmes à +/- 1 près. Cette règle ne sélectionne pas seulement des candidats lexicaux partageant le même début. Ainsi, il semblerait que ce modèle ne permette pas d'expliquer que l'effet d'amorçage puisse corréler avec la densité du voisinage phonologique.

6.1.3.2. Interprétation des résultats des Etudes 3 et 4 pour le modèle NAM

Rappelons que la spécificité majeure du modèle NAM est qu'il prédit que d'une part, la présentation d'un signal acoustique de parole va activer sa représentation en mémoire mais également l'ensemble de ses voisins phonologiques, lors de la phase de génération des candidats lexicaux. Il fait également l'hypothèse que plus un mot a de voisins phonologiques plus le nombre de compétiteurs activés sera important et plus la reconnaissance de ce mot sera gênée par la suite. Le modèle NAM permet donc, comme les autres modèles d'accès au lexique, d'expliquer que la présentation d'un mot entier en tant qu'amorce (e.g., /poze/, « poser ») facilite la reconnaissance ultérieure du mot-cible « poser », par rapport à une condition contrôle (effet d'amorçage par répétition). De plus, ce modèle postule un effet d'amorçage par répétition plus important lorsque la cible est fréquente dans le langage oral et a peu de voisins phonologiques.

Remarquons que les études ayant utilisé le paradigme d'amorçage par répétition avec une amorce auditive (e.g., Kim et al., 2010) ou visuelle (Buchwald et al., 2009) ont trouvé des résultats compatibles avec ces prédictions. Ainsi, si le modèle NAM incluait le signal visuel de parole comme source d'information dans son architecture, il pourrait rendre compte des

résultats de Buchwald et al. (2009). Néanmoins, cela ne lui permettrait pas de rendre compte de l'effet d'amorçage phonologique partiel obtenus avec une amorce auditive (Spinelli et al., 2001, Etude 3) mais également ceux que nous avons obtenus avec une amorce présentée en modalité V (Etude 3 et 4). En effet, les résultats de l'Etude 3 et 4 de ce chapitre ont mis en évidence que la présentation du geste articulatoire pour /by/ accélérait la reconnaissance auditive du mot-cible /byʁo/« bureau ». Selon le modèle NAM, la présentation de la syllabe /by/ devrait activer tous les voisins phonologiques de /by/ à plus ou moins un phonème près. Or, le mot-cible « bureau » n'en fait pas partie puisqu'il ne diffère pas d'un mais de deux phonèmes avec /by/. Ainsi, le modèle NAM prédit que la présentation de /by/ ne devrait pas influencer la reconnaissance ultérieure du mot « bureau ». Il ne peut donc pas prédire l'existence d'un effet d'amorçage partiel comme celui observé dans les Etudes 3 et 4. De plus, les résultats de l'Etude 4 ont mis en évidence que l'effet d'amorçage partiel était plus important pour des mots-cibles de basse fréquence lexicale, ayant une densité de voisinage phonologique élevée. Même si NAM permettait de rendre compte des effets d'amorçage partiels observés dans les Etudes 3 et 4, celui-ci prédirait une influence contraire de la fréquence et de la densité du voisinage phonologique des mots-cibles (cf. Kim et al., 2010 ; Buchwald et al., 2009). Notons néanmoins que l'interprétation de nos données avec NAM reste relativement difficile, ce modèle ayant uniquement été conçu pour pouvoir rendre compte de la reconnaissance de mots monosyllabiques et que nous avons utilisé des stimuli bisyllabiques dans l'ensemble des travaux présenté dans ce manuscrit.

6.1.3.3. Interprétation des résultats des Etudes 3 et 4 pour les modèles TRACE et Shortlist A

Si le modèle TRACE incluait l'information visuelle dans son architecture, celui-ci pourrait également rendre compte de l'effet principal d'amorçage observé dans les Etudes 3 et 4. En effet, TRACE supposerait que la présentation du geste articulatoire pour /by/ permettrait d'activer la représentation du mot « bureau » au niveau lexical. L'intervalle de 50 ms entre la présentation de l'amorce et de la cible étant très court, l'activation reçue par « bureau » n'aurait pas le temps de complètement se dissiper avant la présentation de l'unité lexicale de « bureau». C'est cette activation résiduelle qui permettrait à TRACE d'expliquer cet effet d'amorçage. De plus TRACE postule que les représentations de mots de haute fréquence ont besoin de recevoir moins d'activation que les mots de basse fréquence pour être reconnus. Ceci permet donc au modèle TRACE de prédire que l'effet d'amorçage puisse être modulé en fonction de ce paramètre. Si les effets d'amorçage en modalité V n'ont été

observés que pour les mots rares, cela peut être dû au fait que l'information visuelle, de par sa nature visémique, facilite le processus de reconnaissance de mots lorsque l'accès au lexique est difficile (i.e., pour les mots de basse fréquence). De plus, nous pensons que TRACE pourrait également expliquer le fait que l'effet d'amorçage obtenu corrèle avec la densité du voisinage phonologique. En effet, TRACE ne postule pas, contrairement au modèle de la cohorte, une importance spécifique du début de mot. Dans ce modèle, « burin » et « sureau » constituent l'un comme l'autre des compétiteurs pour la reconnaissance de « bureau ». Le fait que les effets d'amorçage obtenus en modalité V soient plus importants lorsque la cible dispose d'une densité de voisinage phonologique élevée, suggère également que l'information visuelle facilite le processus de reconnaissance de mots lorsque l'accès au lexique est difficile. Ainsi, dans l'hypothèse que TRACE puisse ajouter l'information visuelle dans son architecture, celui-ci nous semblerait plus apte, par rapport au modèle de la Cohorte II ou NAM, à rendre compte de l'ensemble de nos résultats.

Remarquons ici que si le modèle Shortlist A incorporait l'information visuelle comme signal d'entrée dans son architecture, nous pensons que celui-ci serait également à même de rendre compte des effets d'amorçage partiels obtenus pour les Etudes 3 et 4. En effet, lors de la première phase d'activation des candidats lexicaux (i .e., lors de la formation de la « short-list »), ce modèle ne limite pas le nombre de candidats activés au nombre de voisins phonologiques. Ainsi, /by/ peut, selon ce modèle, activer la représentation du mot « bureau » et ainsi faciliter sa reconnaissance ultérieure. De plus, il semblerait que ce modèle puisse, de la même manière que TRACE rendre compte du fait que l'effet d'amorçage est corrélé avec la densité du voisinage phonologique. En effet, ce dernier n'accorde pas de statut spécifique au début de mot et peut rendre compte de l'influence de la densité du voisinage phonologique sur le processus de reconnaissance de mots (voir Norris & McQueen, 2008, pour une simulation sur une version bayésienne de Shortlist A).

En conséquence, il semble que les résultats dans l'Etude 3 et 4 sont plutôt compatibles avec les modèles tels que TRACE et Shortlist A plutôt qu'avec les modèles de la Cohorte II et NAM. Insistons néanmoins sur le fait que cette compatibilité est hypothétique, aucun d'entre n'incluant à ce jour l'information visuelle dans leur architecture.

Nous attirons également l'attention du lecteur sur le fait que nos données suggèrent que l'apport de l'information visuelle dans le processus d'activation des représentations lexicales est plus important lorsque l'accès au lexique est difficile, c'est-à-dire pour

reconnaître des mots peu fréquents et ayant une densité de voisinage phonologique importante. Comment les versions audiovisuelles hypothétiques de TRACE et Shortlist A, qui semblent les plus aptes à rendre compte de nos données, pourraient expliquer ce phénomène ? Rappelons que ces deux modèles ont en commun le fait qu'ils peuvent rendre compte des effets de fréquence lexicale, en supposant un seuil d'activation plus haut (et donc une quantité d'activation nécessaire plus importante) pour reconnaître un mot de basse que de haute fréquence. Egalement, les effets de voisinage phonologique sont expliqués par ces modèles en postulant que de l'inhibition intra-niveau circule entre les candidats lexicaux qui sont activés par l'arrivée d'un signal de parole (dont les voisins phonologiques font partie). De par ce mécanisme, la représentation lexicale correspondant au signal d'entrée va recevoir plus d'inhibition intra-niveau plus elle a de compétiteurs et donc plus elle a de voisins. En conclusion, pour ces deux modèles, un mot peu fréquent et disposant de peu de voisins nécessite de recevoir beaucoup plus d'activation qu'un mot fréquent, ayant une densité de voisinage phonologique faible. Nous pensons que l'information visuelle ne jouerait un rôle notable dans le processus d'activation des représentations lexicales que lorsque la quantité d'activation nécessaire pour reconnaître un mot est importante. Ainsi, voir le geste articulatoire pour le début d'un mot faciliterait majoritairement l'accès aux mots difficiles à reconnaître. Néanmoins, de par sa nature visémique, elle n'impacterait pas ou peu le décodage des mots fréquents et ayant peu de voisins phonologiques, l'information pouvant être extraite du geste articulatoire n'étant pas assez fine pour faciliter de manière significative la reconnaissance de ces mots.

6.1.4. Interprétation des résultats de l'Etude 5 et perspectives

L'objectif de l'Etude 5 était d'étudier le rôle de l'information visuelle dans le processus d'accès au lexique chez l'enfant. Les résultats ont montré que dès l'âge de 5-6 ans, les enfants ont la capacité d'utiliser l'information lexicale mais aussi visuelle pour détecter un phonème dans une situation bruitée. Néanmoins l'absence d'interaction entre ces deux facteurs suggère que les enfants ne peuvent pas, exploiter le signal visuel de parole à un niveau lexical au moins jusqu'à l'âge de 10 ans. Etant donné que l'information visuelle participe à l'activation des unités lexicales chez l'adulte (i.e., Etude 1, 2, 3 et 4), nous devrions observer ce changement lors de l'adolescence. Ainsi reconduire le même type d'étude chez des participants âgés entre 11 et 18 ans devrait nous permettre de déterminer à partir de quel moment, dans le développement de l'individu, l'information visuelle est exploitée à un niveau lexical.

Remarquons néanmoins que l'absence d'interaction entre le statut lexical et la modalité de présentation des stimuli ne nous permet pas de conclure de manière formelle quant au rôle joué par l'information visuelle dans le processus d'accès au lexique chez l'enfant. Une raison qui permettrait d'expliquer le fait que nous n'ayons pas obtenu d'interaction entre l'influence de l'information visuelle et lexicale vient peut-être du paradigme utilisé. En effet, il est possible qu'effectuer cette tâche, avec une information acoustique détériorée, ait constitué une certaine difficulté pour les enfants, notamment pour les plus jeunes (i.e., 5-6 ans). D'autres études reposant sur la perception d'unités plus faciles à détecter pour les enfants que les phonèmes (e.g., tâches de détection de syllabe, Liberman et al., 1974) et ne nécessitant pas que le signal acoustique soit détérioré devront être effectuées afin de clairement statuer sur ce point. Une autre solution serait de mesurer l'impact de facteurs lexicaux tels que la densité du voisinage phonologique sur le décodage de l'information visuelle seule à différents stades de l'enfance. En effet, plusieurs études indiquent que ces paramètres influencent le traitement de l'information auditive chez des enfants de la même manière que chez les adultes (e.g., Coady & Aslin, 2003; Garlock, et al., 2001; Krull, et al., 2010; Metsala, 1997; Wang, et al., 2010). Evaluer si le décodage du signal visuel de parole chez l'enfant est sensible à ce facteur pourrait nous permettre de tester d'une autre manière si cette information est exploitée, au même titre que l'information auditive, au niveau lexical.

6.2. PERSPECTIVES

6.2.1. Quel type d'unité fonctionnelle est impliqué dans l'accès au lexique en modalité visuelle seule?

Les résultats des Etudes 3 et 4 ont montré que le geste articulatoire correspondant aux deux premiers phonèmes d'un mot constituerait suffisamment d'information pour activer sa représentation lexicale en mémoire. Une question qui découle directement de ce constat consiste à se demander quel type d'unité pré-lexicale est impliqué dans ce processus. Or nous avons vu dans le Chapitre 2 que cette problématique est sujette à débat dans la littérature décrivant l'accès au lexique en modalité auditive. En effet, alors que certains pensent que le phonème permettrait d'accéder directement au lexique (McClelland & Elman, 1986; Pisoni & Luce, 1987), d'autres suggèrent que la syllabe (Mehler et al., 1981) ou les unités phonétiques tels que les traits acoustiques (Marslen-Wilson & Warren, 1994 ; Norris et al., 2000) pourraient jouer ce rôle. Dans les Etudes 3 et 4, nous avons pris soin de sélectionner des amorces dont le geste articulatoire correspondait à la fois aux deux premiers phonèmes d'un mot mais également à sa première syllabe. En conséquence, la facilitation observée dans ces

deux études pourrait être due au fait que les deux phonèmes, la syllabe ou les traits articulatoires contenus dans l'amorce ont/a permis, en tant qu'unité(s) fonctionnelle(s), d'activer les représentations de mots contenues dans le lexique.

Ainsi, il serait possible de tester le rôle de la syllabe dans l'accès au lexique en modalité visuelle seule, en adaptant le paradigme de Mehler et al. (1981, cf. section 2.2.3 du Chapitre 2) à l'étude de cette question. En effet, nous pourrions reconduire une expérience similaire aux études 3 et 4 en manipulant également l'alignement syllabique entre l'amorce visuelle (articulation silencieuse) et le mot-cible auditif (condition « alignée » e.g., /na/-/na.vɛ/, « navet » vs. condition « non alignée » e.g., /nav/-/na.vɛ/). Si un effet d'amorçage est uniquement observé pour la condition « alignée », cela signifierait que la syllabe joue un rôle dans le processus d'accès au lexique en modalité visuelle seule, toutes caractéristiques phonémiques et phonétiques étant égales par ailleurs.

Nous pourrions également tester si le trait phonétique/articulatoire ou le phonème permettent de contacter les unités lexicales en modalité visuelle seule. Le trait phonétique/articulatoire nous paraît être l'hypothèse la plus probable puisque, du fait de la nature visémique de l'information visuelle, la mise en action de certains articulateurs n'est pas visible (e.g., la vibration des cordes vocales pour le voisement) et rend la distinction entre deux phonèmes (e.g., /p/ vs. /b/) très difficile voire impossible en modalité visuelle seule. En effet, la présence de l'information visuelle augmente l'intelligibilité des phonèmes en présence de l'information auditive, mais ne permet pas à elle seule d'identifier des phonèmes. En conséquence, il serait envisageable que lors de la perception de la parole, le signal visuel soit tout d'abord décodé en traits articulatoires et contacte ensuite directement les représentations lexicales, sans passer par un niveau d'analyse phonémique intermédiaire. Cette idée serait ainsi compatible avec les modèles d'accès au lexique tels que le modèle de la Cohorte II ou Merge, ces derniers ne postulant pas d'étape entre le stade d'extraction des traits phonétiques et le niveau lexical. Remarquons qu'en 1991, Summerfield a d'ailleurs évoqué idée similaire : « [...] lexical access in lipreading must generally be made directly from the visible pattern of articulation without individual consonants and vowels being identified at an intermediate stage[46] » (p. 119). Une solution permettant de tester cette hypothèse serait d'évaluer si la présentation d'une amorce visuelle partageant plusieurs traits articulatoires mais aucun phonème en commun avec la cible (e.g., /pu/-/byʁo/) facilite ou non sa reconnaissance.

[46] « L'accès au lexique en situation de lecture labiale doit généralement être effectuée directement à partir des patterns d'articulation visibles, sans que les consonnes et les voyelles ne soient identifiées à un stade intermédiaire de traitement »

6.2.2. L'information visuelle dans l'accès au lexique

6.2.2.1. Un rôle anticipateur ?

Dans le Chapitre 1, nous avons vu que dans certaines situations, le geste articulatoire précède temporellement sa conséquence acoustique (Cathiard, 1994 ; Smeele, 1994). Les travaux de Smeele (1994) suggèrent que cette avance temporelle serait présente naturellement à l'initiale d'un mot dont le premier phonème est une occlusive labiale. Or, il a été montré que cette avance temporelle serait exploitée par notre système perceptif afin d'extraire des traits phonétiques/articulatoires (e.g., place d'articulation bilabiale, Smeele, 1994 ; l'arrondissement, Cathiard, 1994) permettant d'anticiper l'arrivée et le traitement du signal acoustique. Pour rendre compte de ce phénomène, certains auteurs ont supposé que l'information visuelle serait décodée par notre système perceptif avant que l'information auditive ne soit disponible afin de *pré-activer* certaines unités abstraites (e.g., phonétiques, phonémiques, Munhall & Tohkura, 1998) permettant de prédire l'arrivée d'un signal acoustique de parole congruent (e.g., Van Wassenhove et al., 2005). Or, nous avons montré dans les Etudes 3 et 4 que les premiers gestes articulatoires correspondant à la production du début d'un mot constituent suffisamment d'information pour contacter les représentations lexicales. L'idée découlant de ces différents résultats est que l'information visuelle ne permettrait pas seulement de pré-activer des représentations pré-lexicales mais également d' « amorcer » les représentations lexicales (Jesse & Massaro, 2010), permettant ensuite de faciliter le traitement de l'information auditive. La Figure 33 représente cette hypothèse pour le modèle de la Cohorte II.

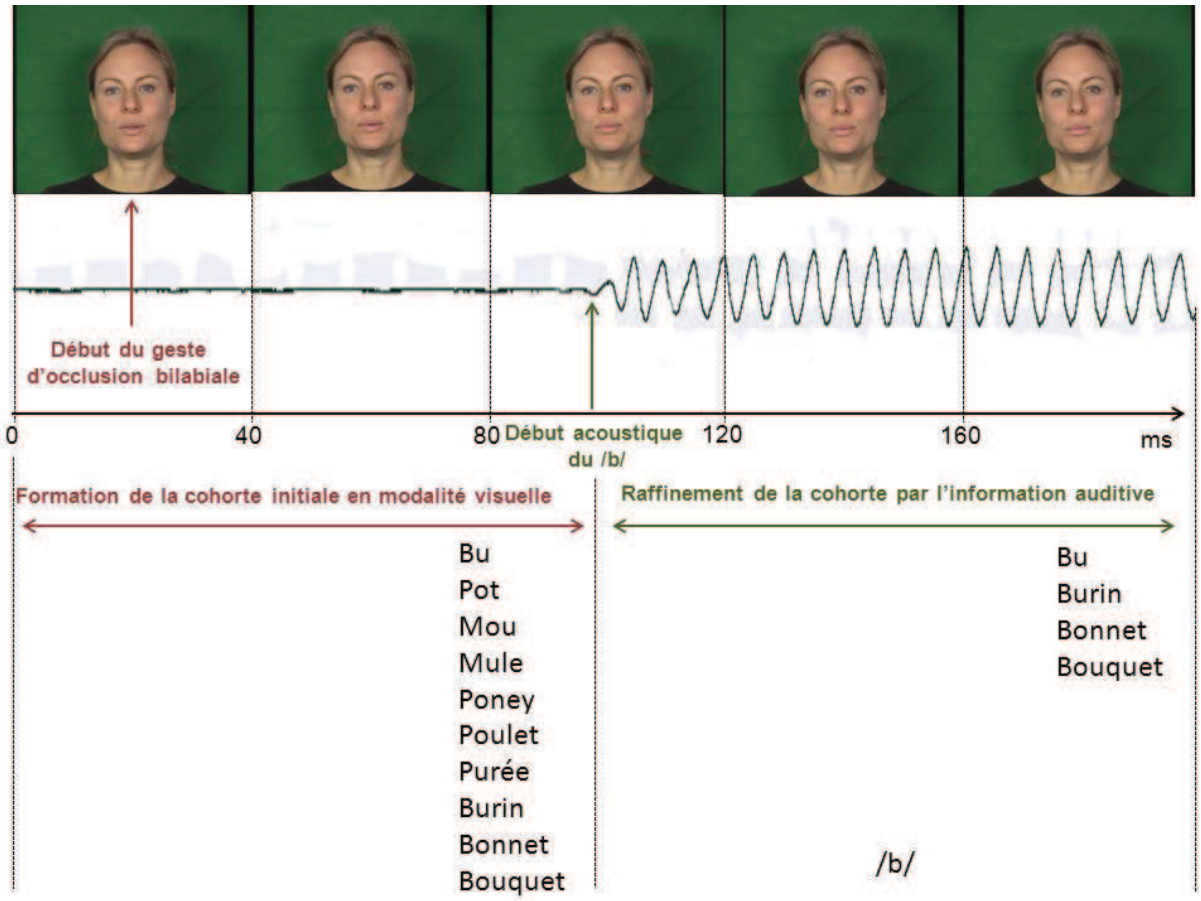


Figure 33. *Représentation schématique de la formation de la cohorte initiale en modalité visuelle et du raffinement de cette cohorte par l'arrivée de l'information auditive, lors de la présentation du début du mot « bonnet ». Notons ici que seuls certains candidats lexicaux ont été sélectionnés pour cet exemple mais que chacune des cohortes devrait contenir un nombre bien plus important de candidats lexicaux.*

Selon cette hypothèse, le fait de voir le geste de fermeture des lèvres précéder l'arrivée du son permettrait de générer une cohorte initiale en modalité visuelle seule avant que toute information auditive ne soit disponible dans le signal. Dans cet exemple, les candidats lexicaux (pré)activés par l'information visuelle permettrait de ne sélectionner que les mots dont le *premier* phonème est consonantique et correspond à un mouvement articulatoire d'occlusion, dont la place d'articulation est labiale. L'arrivée de l'information auditive permettrait de ne garder activés que les candidats lexicaux compatibles avec le signal acoustique. Dans cet exemple, seules les représentations lexicales dont le premier phonème est voisé mais n'est pas nasalisé seront conservées dans la cohorte créée en modalité audiovisuelle. Ainsi, les mots commençant par /b/ font toujours partie de cette cohorte audiovisuelle, alors que ceux commençant par /p/ et /m/ seront exclus. Notons ici que du fait de la nature visémique de l'information visuelle, nous avons supposé que le geste d'occlusion labiale permettrait tout aussi bien de générer des candidats lexicaux commençant par /b/ que par /p/ et /m/, impliquant des cohortes visuelles de taille plus importante que celles effectuées en modalité auditive seule.

Ainsi nous proposons que ce processus de (pré)activation des unités lexicales soit assuré par un mécanisme initial de formation d'une cohorte visuelle dont les hypothèses lexicales seraient raffinées par l'arrivée de l'information auditive. En d'autres termes, nous supposons que l'information visuelle jouerait un rôle facilitateur lors des premières phases de l'accès au lexique. C'est pour cette raison que nous avons délibérément choisi de développer cette idée dans le cadre du modèle de la Cohorte II qui accorde une importance particulière au début de mot lors de la génération des candidats lexicaux. En effet, postuler que la perception d'un geste articulatoire à l'initiale d'un mot permet de prédire sa conséquence acoustique implique que seuls les candidats lexicaux dont le début est compatible avec l'information visuelle vont être activés. En d'autres termes, nous supposons que voir le premier geste articulatoire d'un mot faciliterait uniquement la reconnaissance des mots partageant le même début, mais pas la même fin. Reprendre le paradigme d'amorçage phonologique partiel utilisé pour les études 3 et 4 et manipuler le type de recouvrement entre l'amorce visuelle et la cible (recouvrement initial : /ba/-/bato/, « bateau » vs. recouvrement final : /to/-/bato/, « bateau ») nous permettrait de tester cette hypothèse.

6.2.2.2. Un rôle dans la segmentation de la chaîne parlée ?

Dans ce travail de thèse, nous nous sommes intéressés au rôle de l'information visuelle dans le processus d'accès au lexique, pour la reconnaissance de mots isolés. Une des continuités possibles de nos travaux serait d'étudier l'impact du signal visuel de parole dans le processus d'accès au lexique pour la reconnaissance de mots placés dans le contexte d'une phrase. En effet, comme nous l'avons évoqué dans le Chapitre 2, le signal de parole est continu alors que tout individu adulte en a une perception discrète. Par conséquent, cette opération rajoute une étape de traitement supplémentaire pour tout interlocuteur devant reconnaitre un mot présenté dans une phrase plutôt que de manière isolée. En effet, celui-ci se doit de *segmenter* la chaîne parlée afin de reconnaître les différents mots qui la composent. Alors que cette question a été sujette a de nombreuses études en modalité auditive (voir e.g., Dumay, 2006 ; Shoemaker, 2009, pour des revues), un plus petit nombre de travaux s'est intéressé au rôle du signal visuel de parole dans ce processus de segmentation. En effet, à notre connaissance, seulement quelques travaux se sont penchés sur des questions similaires chez l'adulte (e.g., Basirat, Sato, Schwartz, Kahane, & Lachaux, 2008; Geers, 1978; Sato, Basirat, & Schwartz, 2007; Sell & Kaschak, 2009). Afin de tester cette hypothèse, les travaux effectués par Sato et al. (2007) et par Basirat et al. (2008) ont utilisé l'effet de transformation verbale qui a été pour la première fois décrit en modalité auditive (Warren & Gregory, 1958). Cette illusion renvoie à des changements perceptifs lors de la présentation d'un

stimulus auditif de manière répétitive et continue. Par exemple, lorsqu'un participant écoute une répétition rapide du mot anglais /laɪf/ « life », vie, sa perception bascule du mot « life » au mot /flaɪ/, « fly », mouche, puis de « fly » à « life » et ainsi de suite (voir Basirat, 2010, pour une revue). Une manière de considérer cet effet de transformation verbale est de le voir comme une bascule de la segmentation du signal acoustique. Dans leurs travaux, Sato et al. (2007) et Basirat et al. (2008) ont étudié cet effet de transformation verbale en modalité audiovisuelle avec des stimuli CVCV (/pata/⇔/tapa/). Ces auteurs ont montré que l'ajout d'un visage articulant /pa/ présenté en synchronie avec le signal de /pa/ dans le flux auditif augmentait les transformations verbales bisyllabiques commençant par cette syllabe (i.e., /pata/). Réciproquement, la présentation du geste articulatoire pour /ta/ au même moment que le signal de /ta/ dans le flux auditif générait plus de transformations verbales /tapa/. Ces résultats suggèrent donc que le signal visuel de parole influence le processus de segmentation du signal acoustique. Une des continuités possibles de ces études et de ce travail de thèse serait d'étudier si l'information visuelle facilite également la segmentation de la chaîne parlée pour des énoncés rendus ambigus suite au phénomène de l'élision du schwa (e.g., /pətitru/ « petite roue » vs. « petit trou ») ou de la liaison (e.g., /grãtami/, grand ami vs. grand tamis). Cela nous permettrait d'évaluer si, comme pour l'information auditive (cf. Spinelli, Mcqueen, & Cutler, 2003), certains indices suprasegmentaux (e.g., tels que la durée d'occlusion) peuvent être extraits du signal visuel de parole (voir Geers, 1978, pour une étude chez des individus malentendants) pour distinguer deux hypothèses lexicales.

6.2.3. Décours temporel de l'intégration audiovisuelle de la parole : avant ou après l'accès aux représentations lexicales ?

Comme nous l'avons évoqué précédemment, une conséquence directe de l'incorporation de l'information visuelle à ces modèles consiste à déterminer à quel moment dans le décours temporel de l'accès au lexique, l'information visuelle et auditive sont intégrées l'une à l'autre. Or, nous avons vu dans le Chapitre 1 que cette problématique est au cœur d'un large débat dans la littérature (e.g., Galantucci et al., 2006 ; Massaro & Chen, 2008). L'objectif de cette section consiste à envisager comment les modèles d'accès au lexique pourraient rendre compte de ce phénomène. A notre connaissance, seuls les travaux de Brancazio (1999 ; 2004) ont clairement envisagé d'incorporer l'intégration audiovisuelle de la parole aux modèles d'accès au lexique. Nous tenons à attirer l'attention du lecteur sur le fait que cette section n'a pas pour objectif de prendre parti pour l'une des alternatives décrites ci-dessous, mais consiste simplement à évoquer différentes alternatives ainsi que

leurs conséquences en termes de codage de l'information auditive et visuelle. Notons que nous avons délibérément choisi d'effectuer cette opération pour les modèles TRACE et Merge, ces derniers permettant à la fois rendre compte des résultats de Brancazio (1999 ; 2004) ainsi que des nôtres, en modalité audiovisuelle (Etude 1 et 2, cf. section 6.1.2).

6.2.3.1. Intégration audiovisuelle avant l'accès aux représentations lexicales

6.2.3.1.1. Pour TRACE

La Figure 34 représente plusieurs versions schématiques du modèle TRACE. Le schéma (a) correspond au décodage d'un signal acoustique correspondant à un mot. Les autres schémas (b à d) représentent des versions hypothétiques du modèle TRACE incluant l'information visuelle dans son architecture. Globalement, nous pouvons remarquer que le nombre ou la taille des flèches indiquent que la quantité d'activation reçue par chaque niveau de traitement (et donc le niveau de traitement des phonèmes) est plus importante en modalité audiovisuelle qu'en modalité auditive seule. Comme nous l'avons déjà évoqué dans la section 6.1.2 cela permet d'expliquer le fait (1) qu'un phonème est plus facilement reconnu en présence du signal visuel de parole et (2) que l'influence du lexique sur ce processus est plus importante en modalité audiovisuelle qu'auditive seule (cf. Etude 1 et 2).

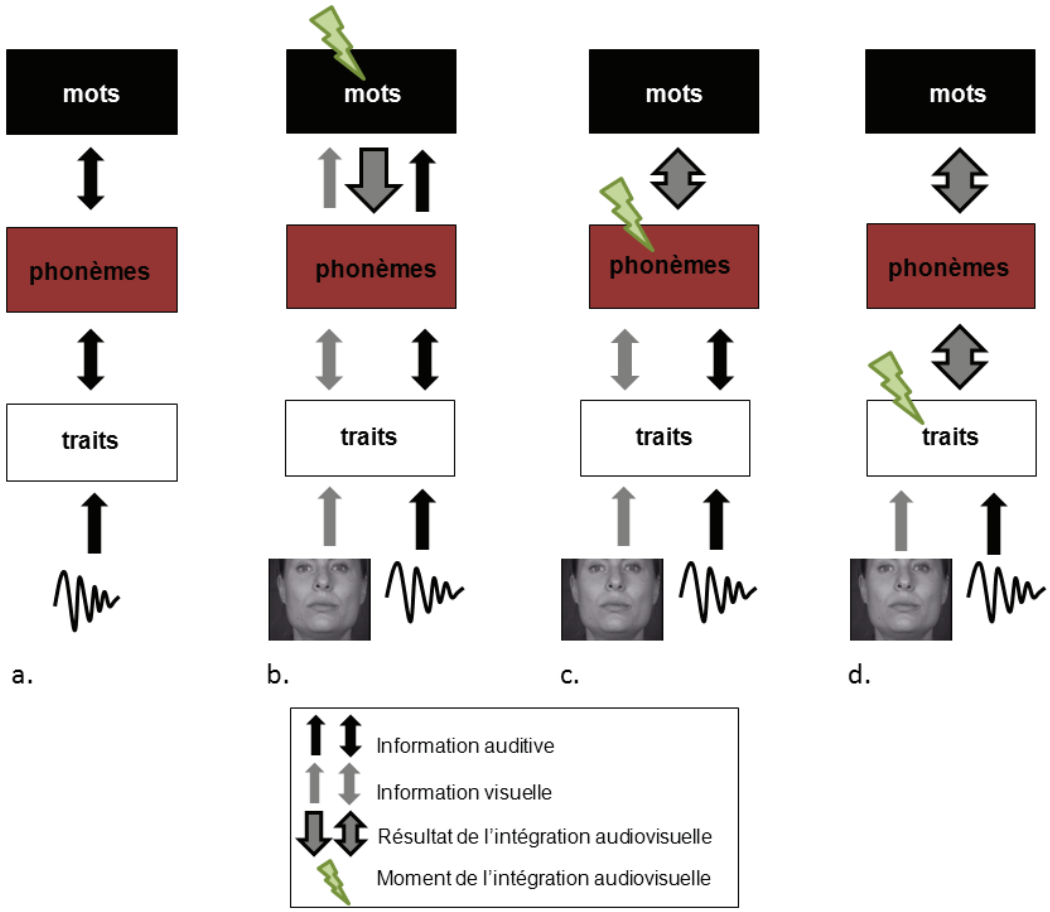


Figure 34. *Représentations schématiques de plusieurs versions de TRACE pour le décodage d'un mot présenté en modalité auditive seule (a) et audiovisuelle (b, c et d). L'intégration audiovisuelle peut s'effectuer au niveau des mots (b), au niveau des phonèmes (c), ou encore au niveau des traits (d). La taille des flèches est proportionnelle à la quantité d'activation. Par souci de simplicité, seules les connexions excitatrices inter-niveaux ont été représentées. La Figure (c) est adaptée de Brancazio (1999).*

Les schémas (b), (c) et (d) permettent de symboliser à quels stades l'information auditive et visuelle pourraient être intégrées dans le cadre du modèle TRACE. Le schéma (b) correspond à une intégration des informations auditives et visuelles à un stade de traitement avancé (i.e., au niveau lexical). Cette configuration suggère que le processus d'intégration des informations auditives et visuelles ne puisse pas être influencé avant le niveau lexical. Par contre, le résultat de l'intégration des informations visuelles et auditives pourrait, par retour d'activation, influencer le niveau des phonèmes. Le schéma (c) représente une version audiovisuelle de TRACE où le processus d'intégration des informations auditives et visuelles s'effectue avant l'activation des représentations lexicales, au niveau des phonèmes. Cette configuration suggère que le résultat de l'intégration des informations auditives et visuelles permet de contacter le lexique. Au niveau des traits cependant, les informations auditives et visuelles sont traitées séparément. Cette configuration est donc la plus compatible avec l'hypothèse de l'intégration « tardive » de Massaro et collègues (Massaro, 1998 ; Massaro & Chen, 2008), qui suggère que les informations visuelles et auditives sont évaluées

séparément pour en extraire des caractéristiques phonétiques/articulatoires et qu'elles sont ensuite fusionnées pour identifier un phonème. Le schéma (d) représente quant à lui une version audiovisuelle du modèle TRACE qui postule que l'intégration audiovisuelle est effectuée dès le premier niveau de traitement, au niveau des traits. Cette version suppose que les informations visuelle et auditive sont fusionnées très tôt dans le processus d'accès au lexique et que seul le résultat de cette intégration audiovisuelle permet de contacter le niveau des phonèmes puis le niveau lexical. Cette configuration nous semble être la plus compatible avec les théories motrices de la perception de la parole (Liberman & Whalen, 2000 ; Galantucci et al., 2006) qui supposent qu'une fusion des informations auditives et visuelles dès les premières phases de la reconnaissance de la parole.

6.2.3.2. Pour Merge

La Figure 35 représente plusieurs versions schématiques du modèle Merge. Le schéma (a) correspond au décodage d'un signal acoustique correspondant à un mot. Les autres schémas (b et c) représentent des versions hypothétiques du modèle Merge incluant l'information visuelle dans son architecture. Globalement, nous pouvons remarquer que de la même manière que pour TRACE, par l'ajout d'une entrée sensorielle, chaque niveau de traitement reçoit plus d'activation (i.e., nombre ou taille de flèches plus importants) en modalité audiovisuelle (schémas (b), (c) et (d)) qu'auditive (schéma (a)), permettant d'expliquer notamment les résultats des Etudes 1 et 2. Le schéma (b) suppose que l'intégration s'effectuerait au niveau lexical. Le niveau de décision phonémique recevrait alors des informations auditives et visuelles séparées du niveau pré-lexical mais également le résultat de l'intégration des informations visuelles et auditives du niveau lexical. Le schéma (c) correspond à une version audiovisuelle de Merge où les informations visuelles et auditives seraient fusionnées dès le niveau pré-lexical. Cette dernière version nous semble la plus compatible avec les théories motrices de la perception de la parole (Liberman & Whalen, 2000 ; Galantucci et al., 2006).

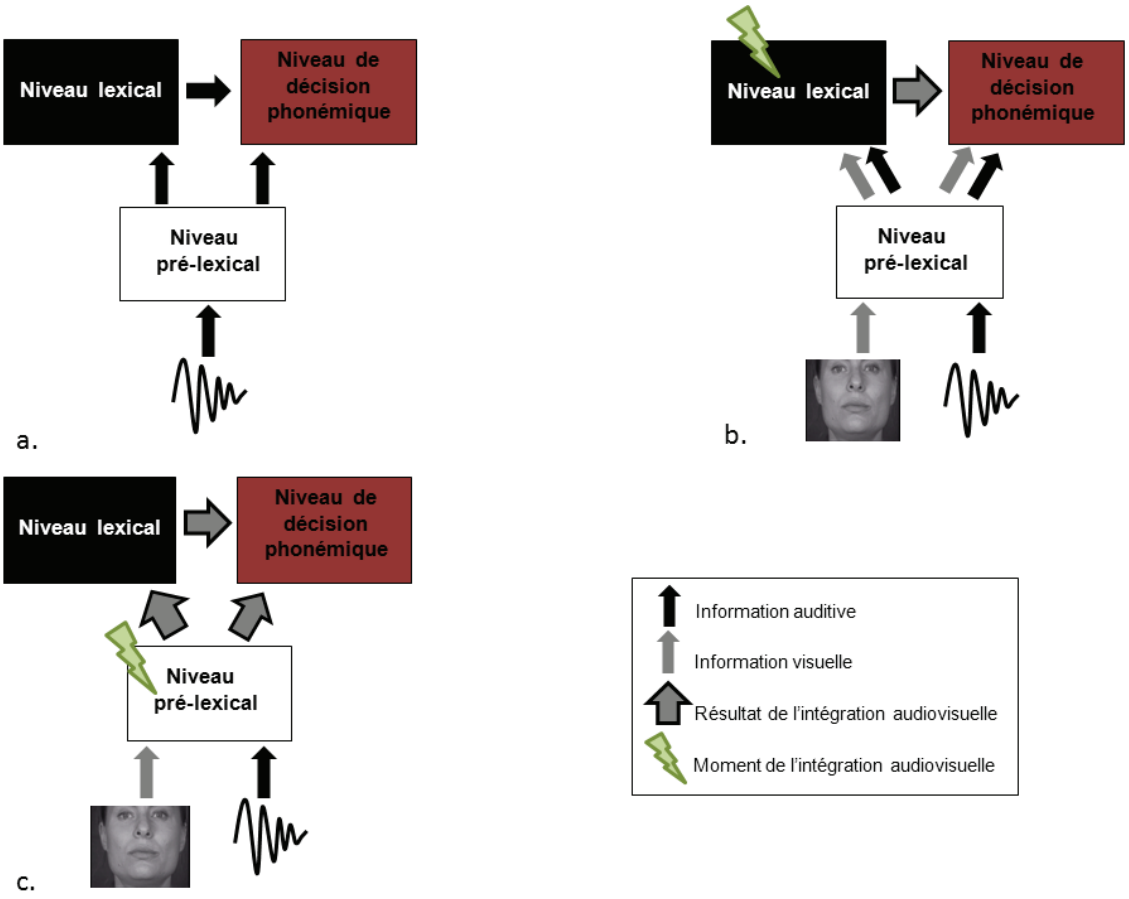


Figure 35. *Représentations schématiques de plusieurs versions de Merge pour le décodage d'un mot présenté en modalité auditive seule (a) et audiovisuelle (b et c). L'intégration des informations auditives et visuelles s'effectue soit au niveau lexical, soit au niveau pré-lexical (c). La taille des flèches est proportionnelle à la quantité d'activation. Par souci de simplicité, seules les connexions excitatrices inter-niveaux ont été représentées. La Figure (b) est inspirée de Brancazio, 1999.*

6.2.3.3. Intégration audiovisuelle après l'accès aux représentations lexicales

La Figure 36 représente des propositions des versions audiovisuelles de TRACE (a) et de Merge (b) supposant que l'intégration des informations auditives et visuelles s'effectuerait après avoir contacté les représentations lexicales. Une conséquence de cette proposition est que les informations auditive et visuelle seraient traitées séparément dans un code spécifique à chaque entrée sensorielle, jusqu'au niveau des mots. Ainsi, ce postulat a pour conséquence que chaque signal est connectée à un lexique différent. En d'autres termes, non pas une mais plusieurs représentations lexicales pourraient permettre de reconnaître un même mot, en fonction de la modalité d'entrée. Notons qu'une idée similaire a été développée dans des modèles récents décrivant l'influence de l'orthographe sur la reconnaissance de mots parlés (e.g., Ferrand, 2001; Grainger, Diependaele, Spinelli, Ferrand, & Farioli, 2003). Cependant, à la différence de l'orthographe, l'information visuelle non seulement influence le décodage de l'information auditive mais est également intégrée à l'information visuelle. Pour

cela, un mécanisme de traduction des informations visuelles et auditives dans un code commun doit être postulé.

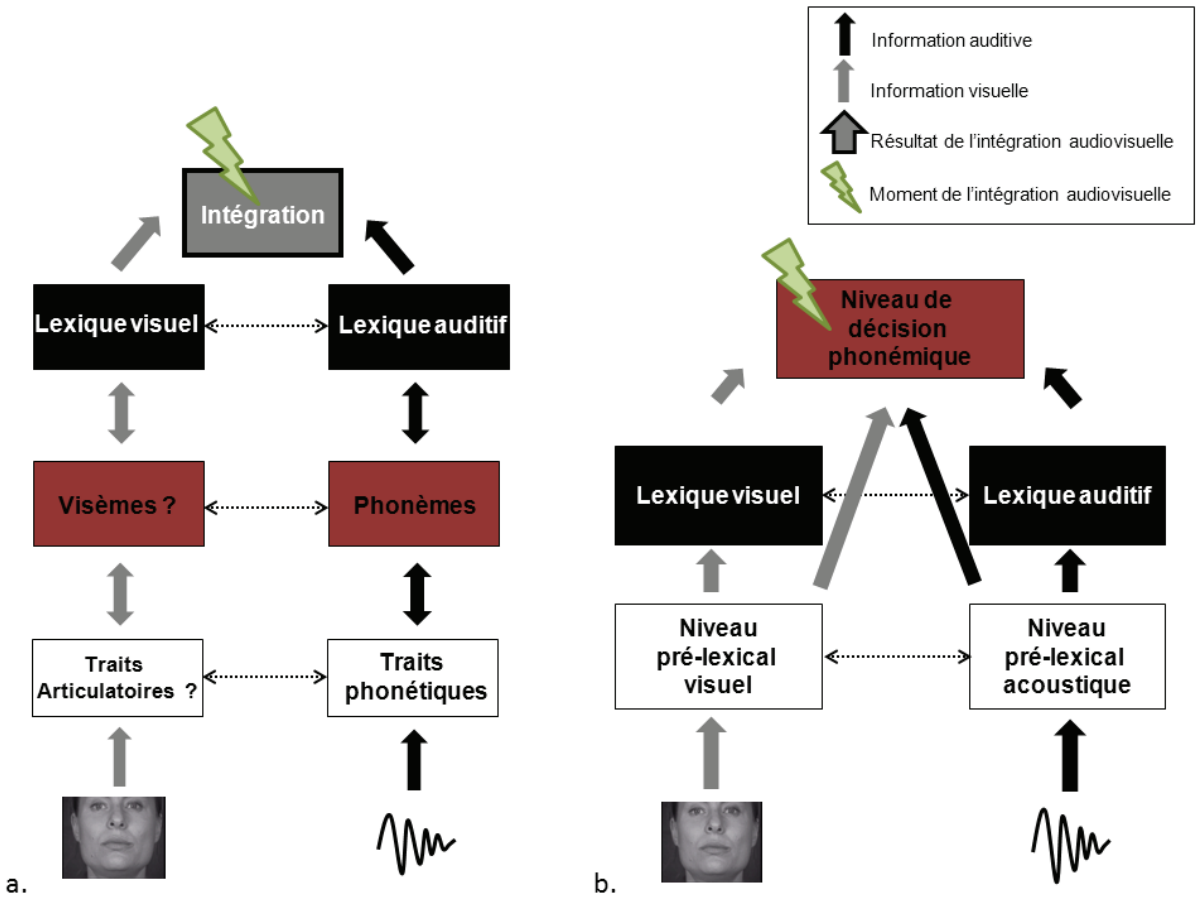


Figure 36. *Représentations schématiques de TRACE (a) et de Merge (b) pour le décodage d'un mot présenté en modalité audiovisuelle, lorsque l'intégration des informations auditives et visuelles s'effectue après l'accès au lexique. La taille des flèches est proportionnelle à la quantité d'activation. Par souci de simplicité, seules les connexions excitatrices inter-niveaux ont été représentées.*

Ainsi, pour TRACE, supposer que les informations visuelles et auditives sont intégrées à un niveau post-lexical nécessite de postuler l'existence d'un niveau supplémentaire de décodage de l'information supplémentaire, permettant de rendre compte de ce phénomène (a). Pour Merge (b), la fusion des informations auditives et visuelles pourraient s'effectuer au niveau de décision phonémique, c'est-à-dire à un niveau de traitement décisionnel. Notons que cette idée est éventuellement compatible avec le modèle à identification directe « Lexical access from Spectra and Faces Parameters » développé par Klatt (1979, cité par Klatt, 1989) Le schéma (d) de la Figure 36 est une représentation de cette hypothèse pour le modèle Merge. Dans les deux cas, des connections inter-niveaux (représentées en pointillés sur le schéma) pourraient rendre compte de l'influence de l'information visuelle sur l'information auditive à différent niveaux, sans pour autant nécessiter un mécanisme de fusion de ces deux informations.

Remarquons néanmoins que cette hypothèse de plusieurs lexiques présente certains inconvénients. En effet, ce type de structure est relativement coûteux en termes d' « espace de stockage » et complexifie l'inclusion de l'information visuelle dans le modèle (ajout de connexions), par rapport aux représentations où un seul et unique lexique est proposé. Cette version du modèle soulève également de nombreuses questions, notamment quant au type de codage et de représentations impliquées dans le traitement spécifique de l'information se propageant par la voie visuelle. De nombreux travaux sont donc nécessaires afin de justifier ou de révoquer l'existence de lexiques séparés pour le traitement de l'information visuelle et auditive.

6.3. CONCLUSIONS

Différentes études réalisées au cours de ce doctorat et présentées dans ce manuscrit nous ont permis de discuter les mécanismes impliqués dans la reconnaissance de mots isolés. Le bénéfice majeur de ces travaux est d'avoir mis en évidence, chez l'adulte, le rôle du signal visuel de parole dans le processus d'activation des candidats lexicaux, en présence ou en l'absence d'information auditive. Egalement, nos travaux indiquent que les premiers gestes articulatoires pour un mot constituent suffisamment d'information pour en activer sa représentation au niveau lexical. Nos résultats obtenus chez l'enfant suggèrent au contraire que jusque l'âge de 10 ans, l'information visuelle ne permet pas encore de contacter les représentations contenues dans le lexique.

En conséquence, les études effectuées chez l'adulte indiquent que les modèles psycholinguistiques actuels décrivant l'accès au lexique devraient inclure la gestualité articulatoire comme source d'information dans leur architecture. Parmi ces modèles, ceux qui semblent le plus pouvoir rendre compte de nos données sont les modèles à activation compétition TRACE (McClelland & Elman, 1986), Shortlist A (Norris et al., 1994) et Merge (Norris et al., 2000).

Un des postulats sous-tendant ce travail de thèse est que chaque individu possède un lexique mental. Nous avons étudié l'influence de ce niveau lexical sur le décodage de la parole audiovisuelle, chez l'adulte et l'enfant d'âge scolaire. C'est dans une perspective plus développementale que nous aimerions étudier, dans le cadre d'un projet postdoctoral, l'émergence et les éventuelles conséquences de l'existence d'un « pseudo lexique » lors de l'apprentissage de mots chez le nourrisson et le jeune enfant. Ce projet serait réalisé sous la direction de Sharon Peperkamp, au Laboratoire de Sciences Cognitives et Psycholinguistique de Paris.

Références

Aboutabit, N. (2007). *Reconnaissance de la Langue Française Parlée Complétée (LPC) : Décodage phonétique des gestes main-lèvres.* Thèse de doctorat. Institut National Polytechnique de Grenoble, France.

Ackroff, J. M. (1981). *The interrelationship of verbal transformations, phonemic restorations and age.* The University of Wisconsin-Milwaukee, Etat-Unis.

Alcorn, S. (1932). The Tadoma Method. *Volta Review, 34*, 195-198.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continous mapping models. *Journal of Memory and Language, 38*, 419-439.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*, 839-843.

Altmann, G. T. M. (1997). Accessing the Mental Lexicon. *The ascent of Babel: An exploration of language, mind, and understanding.* Oxford: Oxford University Press.

Anderson, S., Skoe, E., Chandrasekaran, B., Zecker, S., & Kraus, N. (2010). Brainstem correlates of speech-in-noise perception in children. *Hearing Research, 270*, 151-157.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. New-York: Cambridge University Press.

Arnold, P., & Hill, F. (2001). Bisensory augmentation : A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*, 339-355.

Auer, E. T. (2002). The influence of the lexicon on speech read word recognition: contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin & Review, 9*, 341-347.

Auer, E. T. (2009). Spoken word recognition by eye. *Scandinavian Journal of Psychology, 50*, 419-425.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you 'read' tongue movements ? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication, 52*, 493-503.

Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., & Crewther, D. P. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology, 20*, 1-11.

Basirat, A. (2010). *Émergence des représentations perceptives de la parole: Des transformations verbales sensorielles à des éléments de modélisation computationnelle.* Institut National Polytechnique de Grenoble, Grenoble, France.

Basirat, A., Sato, M., Schwartz, J.-L., Kahane, P., & Lachaux, J.-P. (2008). Parieto-frontal gamma band activity during the perceptual emergence of speech forms. *NeuroImage, 42*, 404-413.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of Neuroscience, 30*, 2414-2417.

Beaud, M. (2006). *L'Art de la thèse...à l'ère du net*. Paris: La Découverte.

Beck, I. M., & McKeown, M. G. (1991). Social studies texts are hard to understand: Mediating some of the difficulties. *Language Arts*, 482-490.

Benguerel, A. P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lipreading. *Journal Of Speech And Hearing Research, 25*, 600-607.

Benoît, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread ? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines*: Berlin: NATO-ASI Series 150 Springer.

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal Of Speech And Hearing Research, 37*, 1195-1203.

Bernstein, L. E. (2005). Phonetic processing by the speech perceiving brain. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 79-98). Oxford: Blackwell.

Bernstein, L. E., Burnham, D., & Schwartz, J.-L. (2002). Special session: Issues in audiovisual spoken language processing (when, where and how?). *7th International Conference on Spoken Language Processing, ICSLP-2002* (pp. 1-4). Denver, Colorado,Etats-Unis.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62*, 233-252.

Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. a. (2009). Development of phonological constancy: toddlers' perception of native- and Jamaican-accented words. *Psychological science, 20*, 539-542.

Binder, J. R., Frost, J. a., Hammeke, T. a., Bellgowan, P. S., Springer, J. a., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10*, 512-528.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal Of Speech And Hearing Research, 17*, 619-630.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16*, 298-304.

Brancazio, L. (1999). *Contribution of the lexicon to audiovisual speech perception.* Thèse de doctorat. University of Connecticut, Etats-Unis.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 445-463.

Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2009). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes, 24*, 580-610.

Burfin, S., Savariaux, C., Granjon, L., Sanchez, C., Tran, T. T. H., Soto-Faraco, S., et al. (2011). Overcoming phonological deafness in L2 conversations by perceiving the facial movements of the speaker. *Workshop on Bilingualism: Neurolinguistic and Psycholinguistic Perspectives.* Aix en Provence, France.

Burnham, D. (1993). Visual recognition of mother by young infants: facilitation by speech. *Perception, 22*(10), 1133-1153.

Burnham, D. (1998). Language specificity in th development of auditory-visual speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by Eye II: Advances in the psychology of speechreading and audiovisual speech*. Hove: Psychology Press.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*, 204-220.

Callan, D. E., Jones, J. a., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*, 2213-2218.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of Auditory Cortex During Silent Lipreading. *Science, 276*, 593-596.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current biology, 10*, 649-657.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research, 12*, 233-243.

Cathiard, M. A. (1994). *La perception visuelle de l'anticipation des gestes vocaliques : cohérence des évènements audibles et visibles dans le flux de la parole.* Thèse de doctorat. Université Stendhal, Grenoble, France.

Cathiard, M. A., Lallouache, M. T., Mohamadi, T., & Abry, C. (1995). Configurational vs. temporal coherence in audio-visual speech perception. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (pp. 218-221). Stockholm: ICPhS.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *Journal of the Acoustical Society of America, 25*, 975-979.

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language, 30*(2), 441-469.

Colin, C., & Radeau, M. (2003). Les illusions McGurk dans la parole : 25 ans de recherches. *L'Année Psychologique, 103*, 497-542.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language, 32*, 193-210.

Connine, C. M., & Clifton, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 291-299.

Connine, C. M., & Titone, D. (1996). Phoneme Monitoring. *Language And Cognitive Processes, 11*, 635-646.

Cox, E. A., Norrix, L., & Green, K. P. (1999). The contribution of visual information to on-line sentence processing : Evidence from phoneme monitoring. *International Conference on Auditory-Visual Speech Processing, AVSP'99*. Santa-Cruz, CA, Etats-Unis.

Cutler, A. (1995). Spoken word recognition and production. In G. A. Miller & P. D. Eimas (Eds.), *Speech, Language and Communication* (pp. 97-136): New-York: Academic Press.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech, 40 (Pt 2)*, 141-201.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*, 385-400.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology, 19*, 141-177.

Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale: Erlbaum.

Cutler, A., & Norris, D. (2002). The role of strong syllables in segmentation for lexical access. In G. T. M. Altmann (Ed.), *Psycholinguistics: Critical concepts in psychology* (pp. 157-177). London: Routledge.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: cross-linguistic comparisons. *Memory & cognition, 28*, 746-755.

Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication, 52*, 555-564.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology, 42*, 317-367.

Dapretto, M., & Bjork, E. L. (2000). The development of word retrieval abilities in the second year and its relation to early vocabulary growth. *Child Development, 71*, 635-648.

Davis, C., & Kim, J. (2001). Repeating and Remembering Foreign Language Words : Implications for Language Teaching Systems. *Artificial Intelligence Review*, 37-47.

Davis, C., Kim, J., & Barbaro, A. (2010). Masked speech priming: neighborhood size matters. *Journal of the Acoustical Society of America, 127*, 2110-2113.

Démonet, J.-F., Thierry, G., & Cardebat, D. (2005). Renewal of the neurophysiology of language: functional neuroimaging. *Physiological Reviews, 85*, 49-95.

Dijkstra, T., Roelofs, A., & Fieuws, S. (1995). Orthographic effects on phoneme monitoring. *Canadian Journal of Experimental Psychology, 49*, 264-271.

Dodd, B., Oerlemens, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli. *Visible Language, 22*, 59-77.

Dufour, S. (2003). *Étude des processus d'activation et de compétition lors de la reconnaissance des mots parlés.* Thèse de doctorat. Université de Bourgogne, Dijon, France.

Dufour, S., & Peereman, R. (2003). Lexical competition in phonological priming: assessing the role of phonological match and mismatch lengths between primes and targets. *Memory & Cognition, 31*, 1271-1283.

Dufour, S., & Peereman, R. (2004). Phonological priming in auditory word recognition: Initial overlap facilitation effect varies as a function of target word frequency. *Current Psychological Letters, 437*.

Dumay, N. (2006). *Rôle des indices acoustico-phonétiques dans la segmentation lexicale: Etudes sur le français.* Thèse de doctorat. Université libre de Bruxelles, Belgique.

Dumay, N., Benraïss, A., Barriol, B., Colin, C., Radeau, M., & Besson, M. (2001). Behavioral and electrophysiological study of phonological priming between bisyllabic spoken words. *Journal of Cognitive Neuroscience, 13*, 121-143.

Dupont, S., Aubin, J., & Ménard, L. (2005). A study of the McGurk effect in 4-and 5-year-old French Canadian children. *ZAS Papers in Linguistics, 40*, 1-17.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal Of Speech And Hearing Research, 12*, 423-425.

Erber, N. P. (1974). Visual perception of speech by deaf children: recent developments and continuing needs. *The Journal of Speech and Hearing Disorders, 39*, 178-185.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience, 15*(2), 399-402.

Feld, J., & Sommers, M. (2011). There Goes the Neighborhood: Lipreading and the Structure of the Mental Lexicon. *Speech Communication, 53*, 220-228.

Ferrand, L. (2001). *Cognition et lecture. Processus de base de la reconnaissance de mots écrits chez l'adulte*. Bruxelles: DeBoeck Université.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal Of Speech And Hearing Research, 11*, 796-804.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 680-698.

Fowler, A. E. (1991). How early phonological development might set the stage for phoneme awareness. In S. Brady & D. Shankweiler (Eds.), *Phonological processes in literacy, A tribute to Isabelle Y Liberman* (pp. 97-117): Lawrence Erlbaum Associates, Inc.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3-28.

Fowler, C. A. (1996). Listener do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99*, 1730-1741.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 816-828.

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 526-540.

Frauenfelder, U. H. (1996). Computational models of spoken word recognition. In T. Dijkstra & K. De Smedt (Eds.), *Computational psycholinguistics: Symbolic and subsymbolic models of language processing.* Londres: Harvester Press.

Frauenfelder, U. H., Scholten, M., & Content, A. (2001). Bottom-up inhibition in lexical selection: Phonological mismatch effects in spoken word recognition. *Language And Cognitive Processes, 16*, 583-607.

Frauenfelder, U. H., Segui, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: facilitatory or inhibitory. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 77-91.

Gagné, J.-P., Rochette, A., & Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Communication, 37*, 213-230.

Gagnepain, P., Chételat, G., Landeau, B., Dayan, J., Eustache, F., & Lebreton, K. (2008). Spoken word memory traces within the human auditory cortex revealed by repetition priming and functional magnetic resonance imaging. *Journal of Neuroscience, 28*, 5281-5289.

Galantucci, B., & Fowler, C. A. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*, 361-377.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110-125.

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-Acquisition, Word Frequency, and Neighborhood Density Effects on Spoken Word Recognition by Children and Adults. *Journal of Memory and Language, 45*, 468-492.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language & Cognitive Processes, 12*, 613-656.

Geers, A. E. (1978). Intonation contour and syntactic structure as predictors of apparent segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 273-283.

Gentil, M. (1981). Etude de la perception de la parole : Lecture labiale et sosies labiaux, . Rapport technique, IBM, France.

Goldinger, S. D. (1996). Auditory Lexical Decision. *Language & Cognitive Processes, 11*, 559-568.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*, 501-518.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F.-H. (2008). Lexical influences on speech perception: a Granger causality analysis of MEG and EEG source estimates. *NeuroImage, 43*, 614-623.

Grainger, J., Bouttevin, S., Truc, C., Bastien, M., & Ziegler, J. (2003). Word superiority, pseudoword superiority, and learning to read: A comparison of dyslexic and normal readers. *Brain and Language, 87*, 432-440.

Grainger, J., Diependaele, K., Spinelli, E., Ferrand, L., & Farioli, F. (2003). Masked repetition and phonological priming within and across modalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1256-1269.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*, 524-536.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*, 267-283.

Grosjean, F., & Frauenfelder, U. H. (1996). A Guide to Spoken Word Recognition Paradigms: Introduction. *Language And Cognitive Processes, 11*, 553-558.

Guiard-Marigny, T., Tsingos, N., Adjoudani, A., Benoît, C., & Gascuel, M.-p. (1996). 3D Models of the Lips for Realistic Speech Animation. *Computer Animation '96* (pp. 80-89). Genève, Suisse.

Hallé, P., & de Boysson-bardies, B. (1994). Emergence of an Early Receptive Lexicon : Infants ' Recognition of Words. *Infant Behavior and Development, 17*, 119-129.

Hallé, P., & de Boysson-bardies, B. (1996). The Format of Representation of Recognized Words in Infants' Early Receptive Lexicon. *Infant Behavior and Development*, 463-481.

Hamburger, M., & Slowiaczek, L. M. (1996). Phonological priming reflects lexical competition. *Psychonomic Bulletin & Review, 3*, 520-525.

Heider, F., & Heider, G. (1940). An experimental investigation of lip-reading. *Psychological Monographs, 52*, 124-153.

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience, 21*, 1229-1243.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*, 393-402.

Hocking, J., & Price, C. J. (2008). The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex, 18*, 2439-2449.

Hockley, N. S., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *Journal of the Acoustical Society of America, 96*, 3309.

Houston, D. (2005). Speech Perception in Infants. In D. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 417-448). Oxford: Blackwell.

Irwin, A. (2008). *Investigating the effect of accent on visual speech.* Thèse de doctorat. Université de Nottingham, Angleterre.

Jackson, P. L., Montgomery, A. A., & Binnie, C. A. (1976). Perceptual dimensions underlying vowellipreading performance. *Journal Of Speech And Hearing Research, 19*, 796-812.

Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: a new multimodal picture-word task. *Journal of Experimental Child Psychology, 102*, 40-59.

Jerger, S., Martin, R. C., & Damian, M. F. (2002). Memory and Language Semantic and phonological influences on picture naming by children and teenagers. *Journal of Memory and Language, 47*, 229-249.

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *72*, 209-225.

Jiang, J. (2003). *Relating Optical Speech to Speech Acoustic and Visual Speech Perception.* Thèse de doctorat. UCLA, Californie, Etats-Unis.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34*, 485-499.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 1-23.

Jutras, B., Gagné, J.-P., Picard, M., & Roy, J. (1998). Identification visuelle et categorisation de consonnes en français quebecois. *Revue d'Orthophonie et d'Audiologie, 22*, 81-87.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language and Hearing Research, 46*, 390-404.

Kandel, S., & Boë, L.-J. (1996). Traitement phonétique et représentation lexicale dans la reconnaissance des mots. *Bulletin de la Communication Parlée*, 61-72.

Kim, J., & Davis, C. (2003). Hearing foreign voices: does knowing what is said affect visual-masked-speech detection? *Perception, 32*, 111-120.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition, 93*, B39-47.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics, 7*, 279-312.

Klatt, D. H. (1989). Reviews of selected model of Speech Perception. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process*. Cambridge: MIT Press.

Klein, E., Moeller, K., Nuerk, H.-C., & Willmes, K. (2010). On the neuro-cognitive foundations of basic auditory number processing: an fMRI study. *Behavioral and Brain Functions, 6*, 42.

Kooijman, V., Hagoort, P., & Cutler, A. (2005). Electrophysiological evidence for prelinguistic infants' word recognition in continuous speech. *Brain research. Cognitive brain research, 24*, 109-116.

Kotz, S. A., Cappa, S. F., Cramon, D. Y., & Friederici, A. D. (2002). Modulation of the Lexical–Semantic Network by Auditory Semantic Priming: An Event-Related Functional MRI Study. *NeuroImage, 17*, 1761-1772.

Krull, V., Choi, S., Kirk, K. I., Prusick, V., & French, B. (2010). Lexical effects on spoken word recognition in children with normal hearing. *Ear and Hearing, 1*, 102-114.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138-1141.

Ladefoged, P. (2001). *Vowels and consonants: an introduction to the sounds of languages*. Oxford: Blackwell.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431-461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4*, 187-196.

Locke, J. L. (1993). *The Child's Path to Spoken Language*. Londres: Harvard University Press.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62*, 615-625.

Luce, P. A., & Lyons, E. A. (1999). Processing lexically embedded spoken words. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 174-183.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing, 19*, 1-36.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Lyxell, B., & Rönnberg, J. (1987). Guessing and speechreading. *British Journal of Audiology, 21*, 13-20.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*, 131-141.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P., et al. (2000). Silent speechreading in the absence of scanner noise: an event-related fMRI study. *NeuroReport, 11*, 1729-1733.

Majerus, S., Collette, F., Van der Linden, M., Peigneux, P., Laureys, S., Delfiore, G., et al. (2002). A PET investigation of lexicality and phonotactic frequency in oral language processing. *Cognitive Neuropsychology, 19*, 343-361.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science, 6*, 314-317.

Mani, N., & Plunkett, K. (2011). Phonological priming and cohort effects in toddlers. *Cognition*, 1-11.

Marchal, A. (2011). *Précis de physiologie de la production de la parole*. Marseille: Solal éditeur.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71-102.

Marslen-Wilson, W. D. (1990). Activation, Competition, and Frequency in Lexical Access. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistics and Computational Perspectives* (pp. 148-172). Cambridge, MA: The MIT Press.

Marslen-Wilson, W. D., Moss, H. E., & Van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1376-1392.

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review, 101*, 653-675.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576-585.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development, 55*, 1777-1788.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge: MIT Press.

Massaro, D. W., & Chen, T. H. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review, 15*, 453-457; discussion 458-462.

Massaro, D. W., Cohen, M. M., & Gesi, A. T. (1993). Long-term training, transfer, and retention in learning to lipread. *Perception & Psychophysics, 53*, 549-562.

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 19-35). Oxford: Oxford University Press.

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41*, 93-113.

Mattys, S. L., Bernstein, L. E., & Auer, E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics, 64*, 667-679.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375-407.

McGrath. (1985). *An examination of cues for visual and audiovisual speech perception using natural and computer generated faces.* Thèse de doctorat. University of Nottingham, Angleterre.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

McQueen, J. M. (2005). Speech perception. In K. Lamberts & R. Goldstone (Eds.), *The Handbook of Cognition* (pp. 255-275). London: Sage Publications.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 37-54). Oxford: Oxford University Press.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1363-1389.

Mehler, J., Dommergues, J., Frauenfelder, U. H., & Segui, J. (1981). The Syllable's Role in Speech Segmentation. *Journal Of Verbal Learning And Verbal Behavior, 20*, 298-305.

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In T. M. Altmann Gerry (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 236-262). Cambridge: MIT Press.

Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition, 1*, 47-56.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338-352.

Möttönen, R. (2004). *Cortical mechanisms of seeing and Hearing speech.* Thèse de doctorat. Helsinki University of Technology, Finlande.

Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of Acoustical Society of America, 104*, 530-539.

Munhall, K. G., & Vatikiotis-bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by Eye II: Advances in the psychology of speechreading and audiovisual speech* (pp. 123-139). Hove: Psychology Press.

Myers, E. B., & Blumstein, S. E. (2008). The neural bases of the lexical effect: an fMRI investigation. *Cerebral Cortex, 18*, 278-288.

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2010). Binding and unbinding in audiovisual speech fusion : Removing the McGurk effect by an incoherent preceding audiovisual context. *International Conference on Auditory-Visual Speech Processing, AVSP2010.* Hakone, Kanagawa, Japon.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*, 4-12.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: a new French lexical database. *Behavior research methods, instruments & computers: a journal of the Psychonomic Society Inc, 36*, 516-524.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du francais contemporain sur internet: LEXIQUE. *L'Année Psychologique, 101*, 447-462.

Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: a retrospective analysis. *Developmental Psychology, 42*, 643-655.

Newman, S. D., & Twieg, D. (2001). Differences in Auditory Processing of Words and Pseudowords : An fMRI Study. *Human Brain Mapping, 47*, 39 - 47.

Nooteboom, N. G., & Doodeman, G. J. N. (1984). Speech quality and the gating paradigm. In M. P. R. van den Broecke & A. Cohen (Eds.), *Proceedings of the I I th International Congress of Phonetic Sciences* (pp. 481-485). Dordrecht, Pays-bas: Foris.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition, 52*, 189-234.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review, 115*, 357-395.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral & Brain Sciences, 23*, 299-325; discussion 325-270.

Ojanen, V. (2005). *Neurocognitive mechanisms of audiovisual speech perception.* Thèse de doctorat. Helsinki University of Technology, Finlande.

Orfanidou, E., Marslen-Wilson, W. D., & Davis, M. H. (2006). Neural response suppression predicts repetition priming of spoken words and pseudowords. *Journal of Cognitive Neuroscience, 18*, 1237-1252.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development, 22*, 237-247.

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*, 93-115.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*, 191-196.

Peperkamp, S., & Dupoux, E. (2002). Coping with phonological variation in early lexical acquisition. In I. Lasser (Ed.), *The Process of Language Acquisition* (pp. 359-385). Francfort: Peter Lang.

Pisoni, D. B., & Levi, S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 3-18). Oxford: Oxford University Press.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition, 25*, 21-52.

Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*, 1037-1052.

Preminger, J. E., Lin, H. B., Payen, M., & Levitt, H. (1998). Selective visual masking in speechreading. *Journal of Speech, Language and Hearing Research, 41*, 564-575.

Pulvermüller, F. (2007). Brain processes of word recognition as revealed by neurophysiological imaging. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 7865-7870.

Radeau, M., Morais, J., & Dewier, A. (1989). Phonological priming in spoken word recognition: task effects. *Memory & Cognition, 17*, 525-535.

Radeau, M., Morais, J., & Segui, J. (1995). Phonological priming between monosyllabic spoken words. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1297-1311.

Raettig, T., & Kotz, S. a. (2008). Auditory processing of different types of pseudo-words: an event-related fMRI study. *NeuroImage, 39*, 1420-1428.

Reed, C. M., Rabinowitz, W. M., Durlach, N. I., Braida, L. D., Conway-Fithian, S., & Schultz, M. C. (1985). Research on the Tadoma method of speech communication. *Journal of the Acoustical Society of America, 77*, 247-257.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of LipReading* (pp. 97-113). Londres: Erlbaum Associates.

Remez, R. E. (1996). Critique: Auditory form and gestural topology in the perception of speech. *Journal of the Acoustical Society of America, 99*, 1695-1698.

Remez, R. E. (2005). The perceptual organization of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 28-50). Oxford: Blackwell.

Repp, B. H., Manuel, S. Y., Liberman, A. M., & Studdert-Kennedy, M. (1983). Exploring the "McGurk effect". *Proceedings of the 24th annual meeting of the Psychonomic Society*. San Diego, Etats-Unis.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169-192.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131-141.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America, 103*, 3677-3689.

Rosenblum, L. D. (2005). Primacy of Multimodal Speech Perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51-78): Oxford: Blackwell.

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon. *Psychological Science, 17*, 405-410.

Rosenblum, L. D. (2010). *See what I'm saying: the extraordinary power of our five senses*. New-York: W.W. Norton & Company.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. *Psychological Science, 18*, 392-396.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*, 347-357.

Ross, L. a., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*, 1147-1153.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 7295-7300.

Rubin, P., Turvey, M. T., & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics, 19*, 384-398.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Sams, M. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication, 26*, 75-87.

Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474-494.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*, 28-51.

Santi, A., Servos, P., Vatikiotis-Bateson, E., Kuratate, T., & Munhall, K. (2003). Perceiving biological motion: dissociating visible speech from walking. *Journal of Cognitive Neuroscience, 15*, 800-809.

Sato, M., Basirat, A., & Schwartz, J.-L. (2007). Visual contribution to the multistable perception of speech. *Perception & Psychophysics, 69*, 1360-1372.

Schwartz, J. L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *The Journal of the Acoustical Society of America, 127*, 1584-1594.

Schwartz, J. L. (2011). Analyse audiovisuelle des scènes de parole. *Rencontres Jeunes Chercheurs en Parole, RJCP2011*. Grenoble, France.

Schwartz, J. L., Basirat, A., Ménard, L., & Sato, M. (2010). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 1-19.

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition, 93*, B69-78.

Schwartz, J. L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by Eye II: Advances in the psychology of speechreading and audiovisual speech* (pp. 85-108). Hove: Psychology Press.

Schwartz, J. L., Sato, M., & Fadiga, L. (2008). The common language of speech perception and action : a neurocognitive perspective. *Perception, 13*, 9-22.

Sebastián-gallés, N., Dupoux, E., Segui, J., & Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language, 31*, 18-32.

Segui, J., Dupoux, E., & Mehler, J. (1990). The Role of the Syllable in Speech Segmentation, Phoneme Identification, and Lexical Access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 263-280). Cambridge: MIT Press.

Seitz, P. F., & Grant, K. W. (1999). Modality, perceptual encoding speed, and time-course of phonetic information. *International Conference on Auditory-Visual Speech Processing, AVSP'99.* Santa-Cruz, CA, Etats-Unis.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science, 11*, 306-320.

Sell, A. J., & Kaschak, M. P. (2009). Does visual speech information affect word segmentation? *Memory & Cognition, 37*, 889-894.

Shahin, A. J., & Miller, L. M. (2009). Multisensory integration enhances phonemic restoration. *Journal of the Acoustical Society of America, 125*, 1744-1750.

Shannon, C. E. (1948). A mathematical theory of communications. *Bell System Technical Journal, 27*, 632-656.

Shoemaker, E. M. (2009). *Acoustic Cues to Speech Segmentation in Spoken French: Native and Non-native Strategies.* Thèse de doctorat. University of Texas, Austin, Texas.

Slowiaczek, L. M., & Hamburger, M. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of experimental psychology Learning memory and cognition, 18*, 1239-1250.

Slowiaczek, L. M., Nusbaum, H. C., & Pisoni, D. B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition, 13*, 64-75.

Smeele, P. (1994). *Perceiving speech: Integrating auditory and visual speech.* Thèse de doctorat. Delft University of Technology, Pays-Bas.

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J. F., & Lewkowicz, D. J. (sous presse). Development of Audiovisual Speech Perception. In A. Bremner, D. Lewkowicz & C. Spence (Eds.), *Multisensory Development*. Oxford: Oxford University Press.

Soto-Faraco, S., Navarra, J., Weikum, W., Vouloumanos, A., Sebastian-Galles, N., & Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception, 69*, 218-231.

Specht, K. (2003). Lexical decision of nonwords and pseudowords in humans: a positron emission tomography study. *Neuroscience Letters, 345*, 177-181.

Spinelli, E. (1999). *Amorçage phonologique et reconnaissance des mots parlés.* Thèse de doctorat. Université Paris Descartes, France.

Spinelli, E., & Ferrand, L. (2005). *Psychologie du langage : l'écrit et le parlé, du signal à la signification*. Paris: Armand Colin.

Spinelli, E., Mcqueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language, 48*, 233-254.

Spinelli, E., Segui, J., & Radeau, M. (2001). Phonological priming in spoken word recognition with bisyllabic targets. *Language and Cognitive Processes, 16*, 367-392.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*(17), 3-45.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of Acoustical Society of America, 111*(4), 1872-1891.

Stevens, K. N. (2005). Features in Speech Perception and Lexical Access. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception*. Oxford: Blackwell.

Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics*(38), 10-19.

Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (2009). Improvement in speech-reading ability by auditory training: Evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. *Neuropsychologia, 47*, 972-979.

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America, 26*, 212-215.

Summerfield, Q. A. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception *Hearing by Eye: The Psychology of LipReading* (pp. 3-51). Londres: Erlbaum Associates.

Summerfield, Q. A. (1991). Visual perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (Vol. 335, pp. 117-137). Hillsdale: Erlbaum Associates.

Swain, I. U., Zelazo, P. R., & Clifton, R. K. (1993). Newborn Infants' Memory for Speech Sounds Retained Over 24 Hours. *Developmental Psychobiology, 2*, 312-323.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 76*, 147-166.

Taft, M., & Hambly, G. (1986). Exploring the Cohort Model of spoken word recognition. *Cognition, 22*, 259-282.

Thomas, S. M., & Jordan, T. R. (2004). Contributions of Oral and Extraoral Facial Movement to Visual and Audiovisual Speech Perception. *30*, 873- 888.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*, 457-472.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology, 12*, 242-248.

Tremblay, C., Champoux, F., Voss, P., Bacon, B. a., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: a developmental study. *PloS one, 2*, e742.

Troille, E., Cathiard, M. A., & Abry, C. (2007). Consequences on bimodal perception of the timing of the consonant an vowel audiovisual flows. *International Conference on Auditory-Visual Speech Processing, AVSP2007*. Kasteel Groenendael, Hilvarenbeek, Pays-Bas.

Trout, J. D., & Poser, W. J. (1990). Auditory and visual influences on phonemic restoration. *Language and Speech, 33 (Pt 2)*, 121-135.

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Development, 75*, 1067-1084.

Tye-Murray, N., & Geers, A. (2001). Children's Audio–Visual Enhancement Test. St. Louis, MO: Central Institute for the Deaf.

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification, 11*, 233-241.

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34*(5), 409-420.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 1181-1186.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*, 926-940.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9*, 325-329.

Vroomen, J., & de Gelder, B. (2004). Perceptual Effects of Cross-modal Stimulation : Ventriloquism and the Freezing Phenomenon. *The Handbook of Multisensory Processes, 3*, 1-23.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal Of Speech And Hearing Research, 20*, 130-145.

Walley, A. C. (1988). Spoken word recognition by young children and adults. *Cognitive Development*, 137-165.

Walley, A. C. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review, 13*, 286-350.

Walley, A. C. (2005). Speech Perception in Childhood. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception*. Oxford: Blackwell.

Wang, N. M., Wu, C. M., & Kirk, K. I. (2010). Lexical effects on spoken word recognition performance among Mandarin-speaking children with normal hearing and cochlear implants. *International Journal of Pediatric Otorhinolaryngology*.

Warren, P., & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics, 41*, 262-275.

Warren, P., & Marslen-Wilson, W. D. (1988). Cues to lexical choice: discriminating place and voice. *Perception & Psychophysics, 43*, 21-30.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392-393.

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology, 71*, 612-613.

Watkins, K. E., Strafella, a. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*, 989-994.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science, 316*, 1159.

Werker, J. F. (2007). The perceptual foundations of phonological development. In G. M. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 579-599). Oxford: Oxford University Press.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701-702.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language, 50*, 212-230.

Woodhouse, L., Hickson, L., & Dodd, B. (2009). Review of visual speech perception by hearing and hearing-impaired people: clinical implications. *International Journal of Language & Communication Disorders, 44*(3), 253-270.

Wright, T. M., Pelphrey, K. a., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex 13*, 1034-1043.

Xiao, Z., Zhang, J. X., Wang, X., Wu, R., Hu, X., Weng, X., et al. (2005). Differential activity in left inferior frontal gyrus for pseudowords and real words: an event-related fMRI study on auditory lexical decision. *Human Brain Mapping, 25*, 212-221.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*, 25-64.

Liste des figures

Liste des tableaux

Annexes

A. MATERIEL UTILISE DANS L'ETUDE 1

Les items expérimentaux (mots et pseudo-mots cibles) utilisés dans l'Etude 1 sont présentés dans le tableau ci-dessous. La fréquence représente la fréquence lexicale en opm. La lettre en gras représente le phonème-cible.

Mots-cibles	Fréquence	Pseudo-mots cibles
A**ff**ût /a**f**y/	1.42	Afé /a**f**e/
A**t**out /a**t**u/	5.74	Ato /a**t**o/
A**v**oue /a**v**u/	61.56	Avo /a**v**o/
Bé**v**ue /be**v**y/	0.36	Bévo /be**v**o/
Bi**s**ou /bi**z**u/	18.4	Bisé /bi**z**e/
Bou**ff**i /bu**f**i/	1.16	Bouffu /bu**f**y/
Ca**d**eau /ka**d**o/	125.79	Cadu /ka**d**y/
Cha**p**eau /ʃa**p**o/	54.91	Chapu /ʃa**p**y/
Choi**s**it /ʃwa**z**i/	170.48	Choisé /ʃwa**z**e/
Cla**p**ot /kla**p**o/	0.16	Clapi /kla**p**i/
Con**f**us /kõ**f**y/	11.02	Confé /kõ**f**e/
Con**v**ie /kõ**v**i/	2.52	Convé /kõ**v**e/
Cou**t**eau /ku**t**o/	58.15	Couti /ku**t**i/
Dé**f**i /de**f**i/	12.24	Défu /de**f**y/
De**v**is /dø**v**i/	0.94	Devé /dø**v**e/
Di**ss**out /di**s**u/	3.94	Dissé /di**s**e/
En**v**ie /ã**v**i/	213.96	Envé /ã**v**e/
Fu**s**il /fy**z**i/	48.61	Fusou /fy**z**u/
Gla**c**is /gla**s**i/	0.02	Glacu /gla**s**y/
In**d**ou /ẽ**d**u/	0.03	Indé /ẽ**d**e/
Lan**d**au /lã**d**o/	1.01	Landi /lã**d**i/
Man**t**eau /mã**t**o/	39.97	Mantu /mã**t**y/
Me**ss**ie /me**s**i/	0.69	Messé /me**s**e/
Na**z**i /na**z**i/	7.11	Nazé /na**z**e/
Ni**v**eau /ni**v**o/	50.7	Nivi /ni**v**i/
Oi**s**eau /wa**z**o/	77.73	Oisi /wa**z**i/
Per**d**u /pɛʁ**d**y/	36.46	Perdo /pɛʁ**d**o/
Pro**f**it /pʁo**f**i/	14.29	Profé /pʁo**f**e/
Ré**c**it /ʁe**s**i/	9.89	Ressé /ʁe**s**e/
Ren**d**u /ʁã**d**y/	508.81	Rendeux /ʁã**d**ø/
So**s**ie /so**z**i/	5.36	Sozou /so**z**u/
Sur**t**out /syʁ**t**u/	1.89	Surti /syʁ**t**i/
Sur**v**ie /syʁ**v**i/	11.64	Survou /syʁ**v**u/
Vau**d**ou /vo**d**u/	2.93	Vaudo /vo**d**o/

B. MATERIEL UTILISE DANS L'ETUDE 2

Les items expérimentaux (mots et pseudo-mots cibles) utilisés dans l'Etude 2 sont présentés dans le tableau ci-dessous. La fréquence lexicale (Freq.) est exprimée en opm. La lettre en gras représente le phonème-cible.

Mots-cibles	Fréquence	Pseudo-mots cibles	Mots-cibles	Fréquence	Pseudo-mots cibles
Bambou /bãb**u**/	1.32	Kambou /kãb**u**/	Mardi /ma**ʁ**d**i**/	22.38	Sardi /sa**ʁ**d**i**/
Bateau /bat**o**/	106.55	Lateau /lat**o**/	Menu /m**œ**n**y**/	9.87	Renu /**ʁœ**n**y**/
Beaucoup /bok**u**/	626	Docoup /dok**u**/	Merci /m**ɛʁ**s**i**/	378.44	Lerci /l**ɛʁ**s**i**/
Bisou /biz**u**/	13.99	Risou /**ʁ**iz**u**/	Monter /m**õ**t**e**/	6.11	Lonter /l**õ**t**e**/
Bureau /by**ʁo**/	156.68	Gureau /gy**ʁo**/	Niveau /niv**o**/	45.46	Tiveau /tiv**o**/
Cadeau /kad**o**/	98.09	Madeau /mad**o**/	Nouveau /nuv**o**/	170.28	Gouveau /guv**o**/
Caillou /kaj**u**/	4.11	Naillou /naj**u**/	Panneau /pan**o**/	9.87	Tanneau /tan**o**/
Casser /kas**e**/	9.05	Dasser /das**e**/	Parler /pa**ʁ**l**e**/	15.82	Darler /da**ʁ**l**e**/
Cerveau /s**ɛʁ**v**o**/	57.67	Perveau /p**ɛʁ**v**o**/	Partout /pa**ʁ**t**u**/	141.94	Nartout /na**ʁ**t**u**/
Chanter /ʃãt**e**/	48.12	Panter /pãt**e**/	Perdu /p**ɛʁ**d**y**/	217.23	Cerdu /m**ɛʁ**d**y**/
Chapeau /ʃap**o**/	48.61	Tapeau /tap**o**/	Pérou /pe**ʁ**u/	0.01	Guerou /ge**ʁ**u/
Choisi /ʃwaz**i**/	58.19	Roisi /**ʁ**waz**i**/	Petit /pət**i**/	573.72	Metit /mət**i**/
Compter /kõt**e**/	31.49	Sonter /sõt**e**/	Plateau /plat**o**/	15.73	Clateau /klat**o**/
Copie /kop**i**/	16.88	Dopie /dop**i**/	Prévu /p**ʁ**ev**y**/	55.54	Crévu /k**ʁ**ev**y**/
Crédit /k**ʁ**ed**i**/	25.82	Brédit /b**ʁ**ed**i**/	Radis /**ʁ**ad**i**/	1.81	Fadis /fad**i**/
Debout /dəb**u**/	91.81	Kebout /kəb**u**/	Reçu /**ʁ**əs**y**/	76.46	Beussu /bəs**y**/
Début /deb**y**/	109.88	Nébut /neb**y**/	Rendu /**ʁ**ãd**y**/	48.31	Lendu /lãd**y**/
Défi /defi/	10.23	Léfi /lefi/	Revue /**ʁ**əv**y**/	7.79	Pevue /pəv**y**/
Dessous /dəs**u**/	18.14	Meussous /məs**u**/	Rideau /**ʁ**id**o**/	10.81	Sideau /sid**o**/
Dessus /dəs**y**/	111.19	Peussus /pəs**y**/	Saler /sal**e**/	9.2	Naler /nal**e**/
Dîner /din**e**/	51.78	Riner /**ʁ**in**e**/	Samedi /samədi/	44.51	Gamedi /gamədi/
Donner /don**e**/	233.3	Lonner /don**e**/	Secoue /səku/	4.77	Beucoue /bəku/
Drapeau /d**ʁ**ap**o**/	14.66	Grapeau /g**ʁ**ap**o**/	Série /se**ʁ**i/	33.34	Térie /te**ʁ**i/
Fermé /f**ɛʁ**m**e**/	13.72	Termer /t**ɛʁ**m**e**/	Soirée /swa**ʁe**/	94.36	Doiree /dwa**ʁe**/
Folie /fol**i**/	49.01	Bolie /bol**i**/	Sonnerie /s**ɔ**n**ʁi**/	6.03	Monnerie /mon**ʁi**/
Fourmi /fu**ʁ**m**i**/	2.78	Dourmi /du**ʁ**m**i**/	Sortie /so**ʁ**t**i**/	42.58	Dortie /do**ʁ**t**i**/
Fusée /fyz**e**/	6	Dusée /dyz**e**/	Souris /su**ʁi**/	21.94	Gouris /gu**ʁi**/
Fusil /fyz**i**/	36.52	Rusi /**ʁ**yz**i**/	Surtout /sy**ʁ**t**u**/	148.66	Durtout /dy**ʁ**t**u**/
Garder /ga**ʁ**d**e**/	21.71	Narder /na**ʁ**d**e**/	Tabou /tab**u**/	0.54	Nabou /nab**u**/
Gâteau /gat**o**/	42.33	Nateau /nat**o**/	Tapis /tap**i**/	20.13	Rapis /**ʁ**ap**i**/
Geler /ʒəl**e**/	3.61	Neler /nəl**e**/	Tenue /tən**y**/	27.1	Penue /pən**y**/
Génie /ʒen**i**/	34.65	Ménie /men**i**/	Tester /t**ɛ**st**e**/	93.68	Mester /m**ɛ**st**e**/
Genou /ʒən**u**/	11.43	Veunou /vən**u**/	Têtu /t**ɛ**t**y**/	6.17	Retu /**ʁɛ**t**y**/
Goûter /gut**e**/	1.84	Nouter /nut**e**/	Tirer /ti**ʁe**/	10.35	Guirer /gi**ʁe**/
Jaloux /ʒal**u**/	29.87	Paloux /pal**u**/	Tissu /tis**y**/	9.21	Bissu /bis**y**/
Jeter /ʒət**e**/	192.18	Deter /ʒət**e**/	Tomber /tõb**e**/	180.25	Romber /**ʁõ**b**e**/
Jeudi /ʒ**œ**d**i**/	24.58	Feudi /f**œ**d**i**/	Tortue /to**ʁ**t**y**/	4.0	Gortue /go**ʁ**t**y**/
Joli /ʒol**i**/	94.55	Roli /**ʁ**ol**i**/	Toupie /tup**i**/	1.5	Foupie /fup**i**/
Jouer /ʒu**e**/	1.9	Pouer /pu**e**/	Tribu /t**ʁi**b**y**/	0.01	Fribu /f**ʁi**b**y**/
Journée /ʒu**ʁ**n**e**/	165.35	Bournée /bu**ʁ**n**e**/	Vaisseau /v**ɛ**s**o**/	67.11	Naisseau /n**ɛ**s**o**/
Jumeau /ʒym**o**/	1.27	Lumeau /lym**o**/	Vaudou /vod**u**/	2.92	Paudou /pod**u**/
Laver /lav**e**/	9.73	Daver /dav**e**/	Vendu /vãd**y**/	37.62	Nendu /nãd**y**/
Lundi /lẽd**i**/	36.01	Vundi /vẽd**i**/	Venue /vən**y**/	93.14	Lenue /lən**y**/
Lycée /lis**e**/	41.96	Rissée /**ʁ**is**e**/	Verrou /ve**ʁ**u/	3.54	Terrou /te**ʁ**u/
Manteau /mãt**o**/	36.16	Ganteau /gãt**o**/	Voler /vol**e**/	3.88	Noler /nol**e**/

C. MATERIEL UTILISE DANS L'ETUDE 3

Les items expérimentaux (Amorces Non Reliées (NR), Amorces Reliées (R) et mots-cibles) utilisés dans l'Etude 3 sont présentés dans le tableau ci-dessous.

Amorces NR	Amorces R	Mots-cibles	Amorces NR	Amorces R	Mots-cibles
mou /mu/	bo /bo/	Beauté /bote/	mu /my/	po /po/	Poney /ponɛ/
vo /vo/		Bottine /botin/	mu /my/		Potier /potje/
vou /vu/		Bonnet /bone/	bu /by/		Polaire /polɛʁ/
vu /vy/	bou /bu/	Bouchon /buʃõ/	su /sy/		Pochette /poʃɛt/
fo /fo/		Bouquet /bukɛ/	su /sy/		Pommier /pomje/
mu /my/		Boulon /bulõ/	vu /vy/		Poème /poɛm/
pu /py/		Bouteille /butɛj/	vu /vy/		Potion /posjõ/
so /so/		Bougie /buʒi/	fu /fy/	pou /pu/	Poumon /pumõ/
so /so/		Boulet /bulɛ/	fu /fy/		Poupée /pupe/
vo /vo/		Bouton /butõ/	fu /fy/		Pouvoir /puvwaʁ/
fo /fo/	bu /by/	Bureau /byʁo/	pu /py/		Poulet /pulɛ/
fou /fu/		Bûcheron /byʃəʁõ/	su /sy/		Poussière /pusjɛʁ/
mou /mu/		Buffet /byfɛ/	su /sy/		Poussin /pusɛ̃/
bou /bu/	fo /fo/	Folie /foli/	vo /vo/		Poubelle /pubɛl/
pou /pu/		Fossile /fosil/	vo /vo/		Poulain /pulɛ̃/
pu /py/		Faucon /fokõ/	vo /vo/		Poussette /pusɛt/
sou /su/		Forêt /foʁɛ/	fou /fu/	pu /py/	Public /pyblik/
su /sy/		Fauteuil /fotœj/	so /so/		Punaise /pynɛz/
bu /by/	fou /fu/	Fourrure /fuʁyʁ/	vou /vu/		Purée /pyre/
su /sy/		Foulard /fulaʁ/	bou /bu/	so /so/	Sauna /sona/
mo /mo/	Fu /fy/	Fusil /fyzi/	bu /by/		Sonnette /sonɛt/
po /po/		Futur /fytyʁ/	fou /fu/		Solide /solid/
pou /pu/		Fumer /fyme/	fu /fy/		Sommeil /somɛj/
so /so/		Fusée /fyze/	mu /my/		Saucisse /sosis/
bu /by/	mo /mo/	Module /modyl/	pu /py/		Sauterelle /sotəʁɛl/
fu /fy/		Modèle /modɛl/	pou /pu/		Soja /soʒa/
fu /fy/		Mollet /molɛ/	vu /vy/		Saumon /somõ/
pu /py/		Motif /motif/	bo /bo/	sou /su/	Soutien /sutjɛ̃/
su /sy/		Moteur /motœʁ/	bu /by/		Souper /supe/
vu /vy/		Moment /momã/	bu /by/		Sourire /suʁiʁ/
vu /vy/		Moral /moʁal/	fou /fu/		Souris /suʁi/
bu /by/	mou /mu/	Moutarde /mutaʁd/	mo /mo/		Souci /susi/
fo /fo/		Mouchoir /muʃwaʁ/	bo /bo/		Soucoupe /sukup/
po /po/		Moulin /mulɛ̃/	mo /mo/		Soudure /sudyʁ/
po /po/		Mouton /mutõ/	pu /py/		Souffrance /sufʁãs/
pu /py/		Moucheron /muʃəʁõ/	vo /vo/		Souplesse /suplɛs/
bou /bu/	mu /my/	Musée /myze/	fou /fu/	su /sy/	Sucette /sysɛt/
po /po/		Muguet /mygɛ/	fou /fu/		Support /sypoʁ/
sou /su/		Musique /muzik/	mo /mo/		Sujet /syʒɛ/
vo /vo/		Museau /myzo/	mo /mo/		Supplice /suplis/
sou /su/	po /po/	Pommade /pomad/	bou /bu/	vo /vo/	Volaille /volaj/
sou /su/		Potage /potaʒ/	bou /bu/		Voleur /volœʁ/
mu /my/		Paupière /popjɛʁ/	sou /su/		Vautour /votuʁ/
mu /my/		Poteau /poto/	sou /su/		Volume /volym/
mu /my/		Police /polis/	su /sy/		Volant /volã/

D. MATERIEL UTILISE DANS L'ETUDE 4

L'ensemble des items expérimentaux utilisé dans l'Etude 4 (Amorces Non Reliées (NR), Amorces Reliées (R) et mots-cible) est répertorié dans le tableau ci-dessous. . La fréquence lexicale (Freq.) est exprimée en opm. La colonne « Nb Voisins » représente la densité du voisinage phonologique.

Amorces NR	Amorces R	Mots-cibles HF	Fréquence	Nb Voisins	Mots-cibles BF	Fréquence	Nb Voisins
le /le/	ba /ba/	Bateau /bato/	106.55	19	Baquet /bakɛ/	0.46	25
re /e/	bo /bo/	Beaucoup /boku/	626	1	Bolet /bolɛ/	0	17
re /e/	bo /bo/	Beauté /bote/	68.57	23	Baudet /bodɛ/	0.11	11
tan /tã/	bu /by/	Bureau /byʁo/	156.68	14	Burin /byʁɛ̃/	0.57	10
vo /vo/	ca /ka/	Cadeau /kado/	98.09	22	Caban /kabã/	0	15
min /mɛ̃/	co /ko/	Côté /kote/	250.51	30	Copeau /kopo/	0.1	12
Pa /pa/	cou /ku/	Couteau /kuto/	51.08	11	Coupon /kupõ/	0.51	17
cou /ku/	de /de/	Début /deby/	109.88	7	Débit /debi/	1.1	14
mu /my/	di /di/	Dîner /dine/	84.73	21	Divan /divã/	1.03	9
je /ʒə/	do /do/	Donner /done/	233	16	Doper /dope/	0.35	11
ri /ʁi/	fa /fa/	Façon /fasõ/	212.6	15	Fagot /fago/	0.03	9
gui /gi/	fo /fo/	Folie /foli/	122.47	8	Fauter /fote/	0	21
guin /gɛ̃/	fu /fu/	Fumer /fume/	35.91	14	Futon /fytõ/	0.28	11
bu /by/	ga /ga/	Gâteau /gato/	42.33	14	Gadoue /gadu/	0.37	7
chu /ʃy/	ma /ma/	Matin /matɛ̃/	265.03	22	Magot /mago/	2.24	11
cou /ku/	mi /mi/	Midi /midi/	35.15	6	Mica /mika/	0.34	12
Kin /kɛ̃/	mo /mo/	Moment /momã/	403	11	Moka /moka/	0.54	12
Kin /kɛ̃/	mo /mo/	Monnaie /monɛ/	26.82	14	Momie /momi/	2.45	11
jou /ʒu/	ni /ni/	Niveau /nivo/	45.46	5	Nicher /niʃe/	0.35	12
li /li/	nou /nu/	Nouveau /nuvo/	170.28	1	Nougat /nuga/	0.89	5
so /so/	pa /pa/	Paquet /pakɛ/	36.9	25	Patin /patɛ̃/	1.12	24
chi /ʃi/	po /po/	Poser /poze/	73.73	21	Potée /pote/	0.04	24
mi /mi/	sa /sa/	Salon /salõ/	37.06	25	Sabot /sabo/	1.79	13
fe /fe/	so /so/	Sauter /sote/	57.89	26	Saumon /somõ/	3.65	11
ro /ʁo/	se /se/	Série /seʁi/	33.34	16	Sénat /sena/	1.38	9
Pe /pə/	su /sy/	Sujet /syʒɛ/	107.92	9	Sumo /symo/	0.88	4
ru /ʁy/	ti /ti/	Tirer /tiʁe/	113.71	31	Titan /titã/	1.06	13
me /me/	tou /tu/	Toucher /tuʃe/	49.46	16	Toupie /tupi/	1.5	5
ru /ʁy/	ve /ve/	Vécu /veky/	51.14	1	Vérin /veʁɛ̃/	0.05	5
ru /ʁy/	ve /ve/	Vélo /velo/	32.95	7	Verrue /veʁy/	0.66	7

E. MATERIEL UTILISE DANS L'ETUDE 5

L'ensemble des items expérimentaux (mots et pseudo-mots cibles) est répertorié dans le tableau ci-dessous. La lettre en gras représente le phonème-cible.

Mot-cible	Pseudo-mot cible
Bateau /bat**o**/	Lateau /lat**o**/
Cadeau /kad**o**/	Madeau /mad**o**/
Casser /kas**e**/	Dasser /das**e**/
Chanter /ʃãt**e**/	Panter /pãt**e**/
Chapeau /ʃap**o**/	Tapeau /tap**o**/
Début /deb**y**/	Nébut /neb**y**/
Dessus /dəs**y**/	Peussus /pəs**y**/
Donner /don**e**/	Lonner /don**e**/
Fermer /fɛʁm**e**/	Termer /tɛʁm**e**/
Garder /gaʁd**e**/	Narder /naʁd**e**/
Gâteau /gat**o**/	Nateau /nat**o**/
Goûter /gut**e**/	Nouter /nut**e**/
Jeter /ʒət**e**/	Deter /ʒət**e**/
Laver /lav**e**/	Daver /dav**e**/
Manteau /mãt**o**/	Ganteau /gãt**o**/
Monter /mõt**e**/	Lonter /lõt**e**/
Parler /paʁl**e**/	Darler /daʁl**e**/
Perdu /pɛʁd**y**/	Serdu /sɛʁd**y**/
Tortue /toʁt**y**/	Gortu /goʁt**y**/
Venue /vən**y**/	Leunu /lən**y**/

F. VALORISATION DE LA THESE

Publications

- **Fort, M.**, Spinelli, E., Savariaux, C. & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication. 52* (6), 525-532.

- **Fort, M**., Kandel, S., Chipot, J., Savariaux, C., Granjon, L. & Spinelli, E. (en révision). Visual speech facilitates the early phases of word recognition: Evidence from fragment priming tasks.

- **Fort, M.,** Spinelli, E., Savariaux, C. & Kandel, S. (en révision). Audiovisual word recognition in children.

Communications

- *Présentations affichées*

  - **Fort, M.**, Chipot, J., Kandel, S., Savariaux, C., Granjon, L., Spinelli, E., (2011). From lips to lexicon: Does visual speech activate lexical representations? *The 10th International Symposium on Psycholinguistics*, 13-16 Avril 2011, San Sebastian, Espagne.

  - **Fort, M.**, Chipot, J., Kandel, S., Savariaux, C., Granjon, L., Spinelli, E., (2010). Can visual information on lip gesture influence lexical processing in speech perception? A phonological priming study. *The 20th International Congress on Acoustics*, 23-27 Août 2010, Sydney, Australie.

  - **Fort, M.**, Spinelli, E., Savariaux, C, Kandel, S. (2010). Examining the contribution of visual information in lexical processing in primary school children: Evidence from vowel detection in noise. *Psycholinguistic approaches to speech recognition in adverse conditions*, 8-10 Mars 2010, Bristol, Royaume-Uni.

  - **Fort, M.**, Spinelli, E., Savariaux, C, Kandel, S. (2009). The contribution of visual information in lexical access: evidence from vowel detection. *Psycholinguistics In Flanders*, 18-19 Mai, 2009, Anvers, Belgique.

  - **Fort, M.**, Kandel, S., Spinelli, E. & Savariaux, C. (2008). Word recognition in audiovisual speech: Preliminary results. *Speech and Face to Face Communication, a workshop dedicated to the memory of Christian Benoît*, 27-29 Octobre, 2008, Grenoble, France.

- *Communications orales*

  - **Fort, M.**, Chipot, J., Kandel, S., Savariaux, C, Granjon, L., Spinelli, E. (2010). From lips to lexicon: Does visual speech activate lexical representations? Seminar at the MARCS Auditory Laboratories, Sydney, Australia.

  - **Fort, M.**, Chipot, J., Kandel, S., Savariaux, C, Granjon, L., Spinelli, E. (2010). La lecture labiale influence-t-elle l'accès au lexique? Seminar at the LPNC, Grenoble, France.

  - **Fort, M.**, Chipot, J., Kandel, S., Savariaux, C, Spinelli, E. (2009). Rôle de l'information visuelle dans l'accès au lexique. *Rencontres Jeunes Chercheurs en Parole*, November 16-18, 2009, Avignon, France.

  - **Fort, M.**, Chipot, J., Kandel, S., Savariaux, C, Spinelli, E. (2009). Le processus de reconnaissance des mots dans la perception audiovisuelle de la parole. Seminar at the GIPSA-lab, Grenoble, France.

G. ARTICLE 1 : FORT, M., SPINELLI, E., SAVARIAUX, C. & KANDEL, S. (2010)

Fort, M., Spinelli, E., Savariaux, C. & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication.* 52 (6), 525-532.

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 525–532
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The word superiority effect in audiovisual speech perception

Mathilde Fort [a,*], Elsa Spinelli [a,b], Christophe Savariaux [c], Sonia Kandel [a,b]

[a] Université Pierre Mendès France, Laboratoire de Psychologie et NeuroCognition (CNRS UMR 5105), BP 47, 38040 Grenoble Cedex 9, France
[b] Institut Universitaire de France, 103, bd Saint-Michel, 75005 Paris, France
[c] Université Stendhal, GIPSA-lab, Dpt. Parole et Cognition (CNRS UMR 5216), BP 25, 38040 Grenoble Cedex 9, France

Received 30 March 2009; received in revised form 29 December 2009; accepted 9 February 2010

Abstract

Seeing the facial gestures of a speaker enhances phonemic identification in noise. The goal of this study was to assess whether the visual information regarding consonant articulation activates lexical representations. We conducted a phoneme monitoring task with word and pseudo-words in audio only (A) and audiovisual (AV) contexts with two levels of white noise masking the acoustic signal. The results confirmed that visual information enhances consonant detection in noisy conditions and also revealed that it accelerates the phoneme detection process. The consonants were detected faster in AV than in A only condition. Furthermore, when the acoustic signal was deteriorated, the consonant phonemes were better recognized when they were embedded in words rather than in pseudo-words in the AV condition. This provides evidence indicating that visual information on phoneme identity can contribute to lexical activation processes during word recognition.

Keywords: Audiovisual speech; Lexical access; Speech perception in noise; Word recognition

1. Introduction

When we speak with someone, most of the time, we are in a face-to-face situation. Except when speaking on the phone or hearing the radio, conversations take place in an audiovisual context. Moreover, the environment in which these conversations take place is often noisy. Several studies have shown that the information on the speaker's orofacial gestures enhances phoneme identification, especially in noisy situations (Benoît et al., 1994; Erber, 1969; Sumby and Pollack, 1954; see Green, 1998 for a review). In French, Benoît et al. (1994) showed that under noisy conditions, consonant and vocalic phonemes embedded in VCVCVC nonsense words were better identified in audiovisual than in auditory only presentations. We may thus assume that visible orofacial gestures boost phonemic units' activation during audiovisual speech perception in a noisy environment. The purpose of the study was to assess whether visual information not only enhances phoneme identification in noise but also contributes to the process of word recognition.

Most of the researches in the field of spoken word recognition studied lexical access in an auditory context (Cutler et al., 1987; Frauenfelder et al., 1990; Ganong, 1980; Samuel, 1981; Warren, 1970). Findings such as the word superiority effect (Cutler et al., 1987), Ganong effect (Ganong, 1980) or phonemic restoration (Samuel, 1981; Warren, 1970), suggest that lexical information influences phoneme perception. For example, with a phoneme monitoring task, Cutler et al. (1987) observed that a consonant (e.g. /b/) was detected faster in a word (e.g. ***b**elle*, i.e. beautiful) than in a pseudo-word (e.g. ***b**erre*). This "word superiority effect" suggests that lexical activation can influence phoneme perception even in situations where the acoustic signal is clear.

* Corresponding author. Address: Université Pierre Mendès France, Laboratoire de Psychologie et NeuroCognition, BP 48, 38040 Grenoble Cedex 9, France. Tel.: +33 4 76 82 56 30; fax: +33 4 76 82 78 34.
E-mail addresses: mathilde.fort@upmf-grenoble.fr (M. Fort), elsa.spinelli@upmf-grenoble.fr (E. Spinelli), christophe.savariaux@gipsa-lab.inpg.fr (C. Savariaux), sonia.kandel@upmf-grenoble.fr (S. Kandel).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.02.005

One of the first studies investigating word recognition processing in an audiovisual context was conducted in Finnish (Sams et al., 1998) with a McGurk paradigm (McGurk and MacDonald, 1976). This effect occurs when an acoustic stimulus /ba/ is presented simultaneously with the articulation of /ga/ in the video signal. Most of the time, it results in the perception of /da/. Numerous studies have replicated these findings, suggesting that during audiovisual speech perception, acoustic and visual signals integrate and may even produce perceptual illusions (see Colin and Radeau, 2003 for a review). On this basis, Sams et al. (1998) displayed an auditory word (e.g. ***p****annu*, stove) simultaneously with another word that was presented visually (e.g. ***k****annu*, pitcher). The audiovisual integration should result in the perception of a pseudo-word (e.g. ***t****annu*). In another condition, the authors displayed an auditory pseudo-word (e.g. ***p****iili*) simultaneously with a visual presentation of another pseudo-word (e.g. ***k****iili*). The audiovisual integration should result in the perception of a word (e.g. ***t****iili*, brick). The results revealed that the McGurk effect was not stronger for word responses than for pseudo-word responses. In other words, there was no word superiority effect. The authors concluded that lexical knowledge did not bias audiovisual speech perception at the stage of phonetic perceptual processing.

Brancazio (2004) argued however that the reason why Sams and colleagues did not observe a lexical effect was, among others, that in their study the stimuli differed as to various parameters other than lexical status. To justify his assessment, Brancazio examined this issue avoiding the potentially confounding variables in Sams et al.'s experiment. He combined the McGurk effect with a Ganong paradigm (Ganong, 1980). In this paradigm the participants had to identify a phoneme /t/ or /d/ that varied along a synthesized t–d continuum. When auditory stimuli in the continuum formed words (e.g. ***t****ask*) and pseudo-words (e.g. ***d****ask*), the proportion of /t/ response was systematically higher than /d/. There was a word superiority effect indicating that phoneme perception was biased in favour of words. In Brancazio's study carried out in English, the participants had to identify /b/ and /d/ in two conditions. In the first one, a word was displayed in the acoustic signal (e.g. ***b****eg*) dubbed into a visual pseudo-word (e.g. ***d****eg*). In the second one, a visual word (e.g. ***d****esk*) was dubbed into an acoustic pseudo-word (e.g. ***b****esk*). The results showed that the lexical bias was stronger in the visual word condition than in the auditory word condition. This suggests that lexical activation not only influences auditory perception but also visual processing during word recognition.

A recent study (Barutchu et al., 2008) provides evidence in line with Brancazio's research. They investigated lexical influences on the McGurk effect in words and pseudo-words. For instance, in the word condition, the auditory word ***b****et* was presented with the visual word ***g****et*. In the pseudo-word condition, a visual pseudo-word ***g****ez* was dubbed into an auditory pseudo-word ***b****ez*. They observed more visual responses – i.e. consistent with the visual signal (***g****et* or ***g****ez*) – for words than for pseudo-words. Consequently, these results also suggest that visual speech processing can be influenced by lexical knowledge.

In sum, Barutchu et al. (2008) and Brancazio (2004) showed that visual information on phoneme identity contributed to lexical access whereas the results of Sams et al. (1998) did not yield any word superiority effect. All these studies used the McGurk paradigm which placed the participants in a situation of perceptual conflict. This may introduce ambiguity in phoneme identification because visual and auditory information are not congruent. Thus, to avoid the conflict between auditory and visual information in our research we examined this issue with another paradigm that is widely used to study the auditory spoken word recognition: the phoneme monitoring task.

To our knowledge, only a few studies have investigated word recognition processes in audiovisual speech perception without using the McGurk paradigm (Buchwald et al., 2009; Kim et al., 2004). First, Kim et al. (2004) used a priming procedure combined with a naming task. They studied whether the presentation of a speaker's orofacial gestures as a prime (visual speech prime without the auditory information) would facilitate the processing of a written target. In a first condition, they displayed word primes in visual speech that was followed by a written target which could be identical (e.g. *back*/*back,* identical condition) or unrelated (e.g. *sharp*/*back,* unrelated condition). In a second condition, they displayed pseudo-word primes in visual speech and the following written target could be identical (e.g. *scay*/*scay*) or unrelated (e.g. *nunth*/*scay*). When the stimuli were words, the authors found a facilitatory priming effect in the identical condition compared to the unrelated condition. They did not observe any facilitatory or inhibitory effect when the stimuli were pseudo-words. With the same paradigm, a recent research reported that participants identified spoken words in noise more accurately when the words were preceded by a visual speech prime of the same word compared with a control condition (Buchwald et al., 2009). They also observed that most of the incorrect responses were phonetically close to the target-words. These findings show that the information in the visual speech prime influences both correct and incorrect identifications. These two studies suggest that the information in the visual speech prime contributes to lexical processing by activating the linguistic forms that match the visual signal. In more general terms – as in Brancazio (2004) and Barutchu et al. (2008) – these studies suggest that visual information contributes to lexical access in audiovisual speech perception.

The purpose of the present research was to examine whether the visual cues that contribute to phoneme identification (Benoît et al., 1994) are also involved in the activation of lexical representations during word recognition process. We also used an experimental paradigm where the visual and auditory information were congruent. The

participants were placed in a noisy environment, which is a situation found in everyday life.

We used a phoneme monitoring task which is a paradigm widely used to investigate the lexical influences on auditory speech processing (Gow, 2003; LoCasto et al., 2007; see also Connine and Titone, 1996 for a review). This task not only provides information on correct identification but also reaction time measures that shed some light on the online process of word recognition.

We conducted a phoneme monitoring task with words and pseudo-words displayed in audiovisual (AV) and auditory alone (A) situations. The stimuli were mixed with noise in the acoustic signal to avoid ceiling effects on correct detection scores. We expected to replicate Benoît et al.'s (1994) results with words: correct responses should be higher in AV than in A especially in noisy conditions. Following the rationale in the word recognition domain we should observe higher scores for words than for pseudo-words (i.e. a word superiority effect). Finally, assuming that the information provided by the speaker's orofacial gestures contributes to the activation of lexical units during word recognition, we predicted that the AV advantage would be higher for words than for pseudo-words.

2. Experiment 1

2.1. Method

2.1.1. Participants

Eighty-one native French speakers ranging in age from 18 to 51 years (mean age = 23 years) participated in the experiment. They all had normal or corrected-to-normal vision and reported no auditory disorders.

2.1.2. Stimuli

The stimulus set was composed of 74 disyllabic word/pseudo-word pairs. Thirty-four pairs were target-present trials (i.e. the target phoneme was in the carrier item, see Appendix) and 40 pairs were target-absent trials. The stimuli in each word/pseudo-word pair were identical except for the last vowel (e.g. /ʃa**p**o/, hat vs. /ʃa**p**y/).

2.1.2.1. Target-present trials. For the 34 pairs of target-present trials (or carrier items), the target phoneme was located at the onset of the second syllable (e.g. the target /**p**/ in /ʃa**p**o/ or /ʃa**p**y/) so that each member of a pair activates the same number of lexical candidates until the target phoneme appears (Marslen-Wilson, 1990). Having the target at the end of the stimulus, instead of the beginning, also increases the probability of observing a strong lexical effect (Frauenfelder et al., 1990). We used seven consonant target phonemes: three labials (/p/, /f/, /v/) and four dentals (/d/, /t/, /s/, /z/). In the carrier item, the following vowel could be either rounded (/o/, /u/, /y/, /œ/) or stretched (/i/, /e/). Half of the word/pseudo-word pairs were contrasting for articulatory gestures. One member of the pair could end in a rounded vowel whereas the other ended in a stretched vowel (e.g. /tʀupo/, flock vs. /tʀupi/). In the other half both members of each pair ended in rounded or stretched vowels (e.g. /ʃapo/ vs. /ʃapy/). The mean word frequency for the carrier words was 45.88 pm (LEXIQUE, New et al., 2001). Half of them were considered as frequent ($F > 10$ occurrences per million) and the other half were not ($F < 10$ occurrences per million).

2.1.2.2. Target-absent trials. The 40 word/pseudo-word pairs of target-absent trials (e.g. /tɔʀty/, turtle vs. /tɔʀti/) were constructed using the same phonemes as described for the carrier items. However, these pairs were always preceded by a non-matching target phoneme (e.g. the target /**p**/ in /tɔʀty/ or /tɔʀti/). The mean word frequency for the target-absent words was 42.99 (New et al., 2001). Half of them were considered as frequent ($F > 10$ occurrences per million) and the other half were not ($F < 10$ occurrences per million).

2.1.2.3. Stimuli recording. The stimuli were recorded in a sound proof room by a trained male native French speaker with a green background. Only the head and top part of the neck of the speaker was visible. He had to start making each utterance with his mouth closed and was instructed to avoid blinking during the stimulus pronunciation. A tri-CCD SONY DXC-990P camera and an AKG C1000S microphone were used to make the recording. The recording was digitalized with the Dps Reality v 3.1.9 software to obtain mpeg video files. The soundtrack extracted from the video was used for the auditory only (A) condition in order to have exactly the same acoustic signal in the A and AV conditions.

Target phonemes were recorded in a sound proof room with a Marantz PMD 670 digital recorder in order to obtain wave files. They were pronounced by a 22-year-old female native French speaker in a schwa context (e.g. target /pə/ for the carrier items /ʃapo/ and /ʃapy/). Thus, two different speakers were chosen to produce the speech material (target phonemes and carrier items) to make sure that phoneme detection would not be due to speaker specificity.

We used Matlab 7.1 software to generate the noise and to add it to each spoken utterance. We used two Signal to Noise Ratios[1]: −9 dB vs. −18 dB. As each utterance energy was dependent on its vowel and consonant type (e.g., plosive, fricative) we calculated the mean strength for each stimulus and then added white noise so that the stimuli could have the same Signal to Noise Ratio throughout the whole presentation.

The stimuli were spread out over four experimental lists corresponding to the four presentation conditions:

[1] The Signal to Noise Ratio, often written S/N or SNR, is a measure of signal strength related to background noise. The ratio is usually measured in decibels (dB). We used the following formula: $\mathrm{SNR} = 20\log_{10}(V_s/V_n)$ in which V_s and V_n are respectively the original signal amplitude and the noise amplitude.

A −9 dB; A −18 dB; AV −9 dB; AV −18 dB. Each list contained 8 or 9 pairs of target-present trials and 10 pairs of target-absent trials. Consequently, each target-present or target-absent trial was presented only once to each participant. The presentation condition of each list was counterbalanced between the participants.

2.1.3. Procedure

The participants were tested individually. They sat at 50 cm from a LCD screen (Neovo 17 X-17A) in a darkened sound proof room. Video stimuli were presented at 25 frames/s. The auditory component of the stimuli was provided at a 44,100 Hz sampling rate by two SONY SRS-88 speakers located on both sides of the screen. The experiment was performed using E-Prime 2.0 software (Psychological Software Tools, Pittsburgh, PA). The participants were given a set of oral instructions explaining that they would first hear a consonant target phoneme and then a word or a pseudo-word (carrier) in which they had to detect this target. They were told that the target phoneme could or could not be in the carrier utterance. A Go/No Go response task was employed: participants had to press the space bar of a keyboard as quickly as possible when they heard the target phoneme in the carrier item and do nothing if they did not hear it, using only one hand to give their answer. The participants were told to detect the consonant target phoneme regardless of its orthographic representation.[2]

For each participant, half of the carrier items appeared in AV with the video of the speaker moving (half at −9 dB, half at −18 dB). In the other half (the A condition) the stimuli were only displayed auditorily (half at −9 dB, half at −18 dB). The experiment was divided into two blocks, namely A and AV. The block order was counterbalanced across participants. Between each block, a black screen informed the participants of a change in the presentation modality. For the AV condition participants were told to watch and listen carefully to the stimuli in order to avoid focusing on one modality more than another (cf. Amano and Sekiyama, 1998; Tiippana et al., 2004). Within each block, the participants perceived the first part of the stimuli in one SNR condition (e.g. −9 dB) and the second part in the other SNR condition (e.g. −18 dB). The order of each SNR condition was counterbalanced across participants. Moreover, half of the items within the four conditions contained the target phoneme (target-present trials) and half did not (target-absent trials). Within each condition, the order of the stimuli was randomised. A 10 stimuli-long training session preceded the test.

2.2. Results and discussion

Mean response latencies and percentages of correct phoneme detection were calculated for each participant and each item pair. Two participants were removed from the analyses because they did not respond in the A condition at −18 dB. A 2 (modality: A vs. AV) × 2 (lexical status: word vs. pseudo-word) × 2 (Signal to Noise Ratio: −9 dB vs. −18 dB) within participants ANOVA was conducted by both participants (*F1*) and items (*F2*). We discarded from the analyses the data that were for every condition above or below two standard-deviations (SD) from the mean (2.3% of the data).

2.2.1. Percentage of correct phoneme detection

Table 1 presents the percentage of correct phoneme detection for words and pseudo-words in A and AV for the two noise conditions (−9 dB vs. −18 dB).[3] The analyses revealed a strong AV advantage, $F1$ $(1, 78) = 180.87$, $p < .001$; $F2$ $(1, 33) = 46.68$, $p < .001$. The analyses also showed that the scores were higher at −9 dB than at −18 dB, $F1$ $(1, 78) = 199.34$, $p < .001$; $F2$ $(1, 33) = 67.54$, $p < .001$. These results replicate Benoît et al.'s (1994) findings. The performance was also enhanced when the target phonemes were embedded in words, $F1$ $(1, 78) = 8.23$, $p < .005$; $F2$ $(1, 33) = 3.06$, $p < .01$.

The interaction between lexical status and modality was significant, $F1$ $(1, 78) = 10.34$, $p < .005$; $F2$ $(1, 33) = 6.96$, $p < .05$. The word advantage for phoneme detection was greater in AV than in A, $F1$ $(1, 78) = 23.83$, $p < .001$; $F2$ $(1, 33) = 10.21$, $p = .01$ and both $Fs < 1$, respectively. There was also a significant interaction between lexical status and noise, $F1$ $(1, 78) = 6.95$, $p = .01$; $F2$ $(1, 33) = 4.69$, $p < .05$. In AV, planned comparisons showed that the word superiority effect was higher at −18 dB than at −9 dB, $F1$ $(1, 78) = 24.72$, $p < .001$; $F2$ $(1, 33) = 11.48$, $p < .01$ and $F1$ $(1, 78) = 4.37$, $p < .05$; $F2$ $(1, 33) = 2.54$, $p = .12$, respectively.

2.2.2. Response latencies

Table 2 presents the response latencies for words and pseudo-words in A and AV for the two noise conditions (−9 dB vs. −18 dB).

[2] In French, for instance, the phoneme /f/ can be written "f" or "ph".

[3] To make sure that the participants did not develop any response strategies, we also computed a d' for each stimulus pair, using this formula $d' = z\,(\mathrm{CD}) - z\,(\mathrm{FA})$ in which z represents the inverse of the normal cumulative distribution and CD and FA refers respectively to the mean probability of correct phoneme detection and false alarms. A 2 (modality: A vs. AV) × 2 (Signal to Noise Ratio: −9 dB vs. −18 dB) × 2 (lexical status: word vs. pseudo-word) within participants ANOVA was conducted by participants. We replicated the results obtained by analysing only the correct detection scores. The analyses on d' revealed a strong AV advantage, F $(1, 78) = 212.7$, $p < .001$. The scores were also higher at −9 dB than at −18 dB, F $(1, 78) = 119.3$, $p < .001$. There was a word superiority effect, F $(1, 78) = 10.7$, $p < .005$. The interaction between lexical status and modality was significant, F $(1, 78) = 5.5$, $p < .05$. Planned comparisons revealed that the word advantage was greater in AV than in A at −9 dB, $F\,(1, 78) = 4.91$, $p < .05$ in AV vs. $F\,(1, 78) < 1$ in A; and at −18 dB, $F\,(1, 78) = 6.2$, $p = .01$ in AV vs. $F\,(1, 78) < 1$ in A. There was also a significant interaction between modality and noise, F $(1, 78) = 10.1$, $p < .005$, suggesting that the AV advantage for phoneme detection was greater at −18 dB than at −9 dB.

Table 1
Percentage of correct phoneme detection as a function of modality, Signal to Noise Ratio (in Decibels, dB) and lexical status, in Experiment 1. Numbers in parentheses represent the standard deviation.

Modality of presentation	Words	Pseudo-words	Word superiority effect
−9 dB			
Audio alone	68.2 (20.4)	69.4 (19.26)	−1.2
Audiovisual	90.9 (10.54)	88.1 (11.9)	2.8*
−18 dB			
Audio alone	50.1 (17.38)	49.3 (18.16)	0.8
Audiovisual	76.1 (19.13)	65.3 (18.32)	10.8**

* Word superiority effect significant by participants.
** Word superiority effect significant by participants and by items.

Table 2
Mean response latencies (in ms), as a function of modality, Signal to Noise Ratio (in Decibels, dB) and lexical status in Experiment 1. Numbers in parentheses represent the standard deviation.

Modality of presentation	Words	Pseudo-words	Word superiority effect
−9 dB			
Audio alone	808.3 (108.4)	808.9 (204.5)	0.6
Audiovisual	671.3 (121)	673.8 (95.5)	2.5
−18 dB			
Audio alone	844.6 (244.1)	899.2 (175.8)	54.6
Audiovisual	734.9 (113.7)	759.1 (138.2)	24.2

The analyses revealed a significant main modality effect in favour of the AV condition, $F1$ (1, 78) = 32.27, $p < .001$; $F2$ (1, 33) = 62.32, $p < .001$. There was a significant main effect of the Signal to Noise Ratio, $F1$ (1, 78) = 31.01, $p < .001$; $F2$ (1, 33) = 8.51, $p < .01$. The participants were faster at detecting a consonant phoneme at −9 dB than at −18 dB. Contrary to our expectations, the lexical effect was not significant, $F1$ (1, 78) = 1.02, $p = .27$; $F2$ (1, 33) = 1.44, $p = .23$. No interaction was significant, all $F1$ (1, 78) < 1.

In sum, Experiment 1 revealed that the participants were faster and had higher scores in AV than in A only conditions. They were also faster and performed better when the Signal to Noise Ratio was at −9 dB than at −18 dB. This is in line with Benoît et al.'s (1994) study conducted with non-word stimuli. More interesting for the purpose of our study was that the scores were higher when the participants had to detect the consonant phonemes embedded in words than in pseudo-words. This word superiority effect was even stronger in the AV condition. It should be pointed out however that we observed the word superiority effect only for correct phoneme detection and not for latencies. Moreover, we were not able to replicate the lexical effect in the auditory only condition, as observed in many studies on word recognition (e.g. Cutler et al., 1987). We thus re-conducted Experiment 1 in a non-noisy environment.

3. Experiment 2

3.1. Method

3.1.1. Participants

Thirty-seven native French speakers ranging in age from 18 to 32 years with a mean age of 21.8 years participated in the experiment. They all had normal or corrected-to-normal vision and reported no auditory disorders.

3.1.2. Stimuli and procedure

They were the same as in Experiment 1 but without noise.

3.2. Results

Mean correct phoneme detection percentages and response latencies were calculated for each participant and for each item pair. A 2 (modality: A vs. AV) × 2 (lexical status: word vs. pseudo-word) within participants ANOVA was conducted by participants ($F1$) and items ($F2$). We discarded from the analyses 1% of our data that was above or below two standard-deviations (SD) from the mean. Table 3 presents the response latencies for words and pseudo-words in the A and AV conditions.

The analyses conducted on response latencies revealed a main lexical effect, $F1$ (1, 36) = 8.21, $p < .01$; $F2$ (1, 33) = 11.48, $p < .01$. As in other studies using a phoneme monitoring task, the participants were faster at detecting the target phonemes in words than in pseudo-words. However, we neither obtained a main modality effect nor an interaction between the two factors (all $Fs < 1$). The analyses on correct phoneme detection[4] did not yield any significant effect (all $Fs < 1$).

4. General discussion

The goal of this study was to show that the visual information provided by the speaker's articulatory gestures contributes to lexical activation during word recognition. We conducted a phoneme monitoring task with words and pseudo-words in Audio only (A) and Audiovisual (AV) contexts with two levels of noise masking the acoustic signal (Experiment 1) and without noise (Experiment 2).

Our results replicated Benoît et al.'s (1994) findings. Phoneme detection scores were higher in AV than in A, especially in noisy conditions (Experiment 1). The audiovisual benefit could be explained by the fact that under

[4] We also computed a d' for each stimulus pair and conducted a 2 (modality: A vs. AV) × 2 (lexical status: word vs. pseudo-word) ANOVA by participants. As for the correct phoneme detection, neither main effects nor interaction between the two factors were significant (all $Fs < 1$).

Table 3
Percentage of correct phoneme detection (CR, in %), and mean Response latencies (RT, in ms) as a function of modality and lexical status in Experiment 2. Numbers in parentheses represent the standard deviation.

Modality of presentation	Words	Pseudo-words	Word superiority effect
RT			
Audio alone	478 (93)	491 (98.8)	13
Audiovisual	475 (128)	493 (125)	18
Mean	476	492	16**
CR			
Audio alone	95.1 (7.1)	96.1 (7.4)	−1
Audiovisual	95.9 (7)	94.4 (8.31)	1.5
Mean	94.8	94.7	0.1

** Word superiority effect significant by participants and by items.

deteriorated acoustic conditions, visual and acoustic signals complement each other (Summerfield, 1987). The auditory information (e.g. place of articulation) that has been masked by the noise is available in the visual signal and can be recovered by seeing the lips, teeth, tongue and jaw movements (Miller and Nicely, 1955; Robert-Ribes et al., 1998). The data on latencies indicate that phoneme detection was faster in AV than in A when the acoustic signal is deteriorated. This suggests that the information on the speaker's orofacial gestures not only enhances phoneme identification in noise (Benoît et al., 1994), but it also accelerates phoneme detection.

The results also revealed a main lexical effect. Consonant phonemes were detected better when they were embedded in words rather than in pseudo-words. This word superiority effect indicates that phoneme detection can be influenced by lexical knowledge. In the noisy situation (Experiment 1), the lexical effect was mostly present in the AV condition. This suggests that the lexical effect is not due to auditory information only. Indeed, these results indicate that the presence of visual information not only facilitates phoneme detection but also contributes in the process of word recognition, especially when the auditory information is deteriorated. Our results are in line with Brancazio (2004) and Barutchu et al. (2008) and provide complementary data indicating that the processing of facial information accelerates phoneme perception and enhances lexical activation in a noisy environment. In Experiment 1, we did not observe a lexical effect in the A modality. We do not have a plausible explanation for this lack of results.

In the without-noise conditions (Experiment 2), consonant phonemes were detected faster when they were embedded in words than in pseudo-words. We observed a main lexical effect on response latencies but no significant interaction with the presentation modality. The lack of effect was essentially due to ceiling effects in both A and AV. Indeed, when the conditions for speech perception were optimal (i.e. when the acoustic signal was clear) the auditory information was enough for recognizing the words efficiently (see Spinelli and Ferrand, 2005, for a review on auditory word recognition studies). Further research should be carried out to determine whether the visual information enhances the lexical activation in every face-to-face situation or only when the auditory information is deteriorated or unavailable.

Models of spoken word recognition such as TRACE (McClelland and Elman, 1986) or MERGE (Norris et al., 2000) describe lexical access in the auditory modality only. However, our data showed a word superiority effect in the AV modality, suggesting that visual information plays a role in lexical access. None of these models incorporate the orofacial gestures as a source of information in their architectures. How could models like TRACE and MERGE account for our results if they included visual information?

TRACE assumes that during the perception of an isolated utterance (i.e. a word or a pseudo-word) some activation first spreads from the sensory input to the featural level. Next, the activation flows to the phonemic stage, where the phonemic decisions are made. When the stimulus is a word, the activation scattering then spreads to the lexical level. To account for the word superiority effect on phoneme detection, TRACE assumes that the activation which reaches the lexical stage flows back to the phonemic level. In this model, the activation can spread bidirectionally between the pre-lexical (featural or phonemic) levels and the lexical units. Thus, the phonemic level receives more activation for a phoneme embedded in a word than for a phoneme embedded in a pseudo-word. Consequently, the word superiority effect observed in our experiment would result – according to a top-down view of spoken word recognition – from a feedback from high-level lexical representations to low-level phonemic units. In the AV condition, pre-lexical units would receive activation from the visual and the auditory inputs whereas in the A modality, the activation flow would only emerge from the auditory input. According to this hypothesis, the phonemic stage would receive more activation in AV than in A. This mechanism could explain why visual information enhances and accelerates the phoneme detection process.

MERGE differs from TRACE with respect to the direction of the activation flow between the pre-lexical and lexical stages. MERGE assumes that activation spreads unidirectionally from pre-lexical to lexical nodes. There is no feedback from lexical to pre-lexical stages. To account for the word superiority effects, MERGE integrates a phoneme decision stage that is independent of the other two nodes. This level is entirely devoted to phonemic decision processes and it is not permanently connected to the other nodes. The connections from the lexical nodes to the phoneme decision nodes are operational only when the listener has to make phonemic decisions (e.g. during a phoneme monitoring task). According to MERGE, during the perception of an isolated word, the activation would spread from the input (or pre-lexical) nodes to the lexical nodes and to the

phoneme decision nodes simultaneously. Then, activation flows from the lexical nodes to the phoneme decision nodes. This excitatory connection between the lexical and phoneme decision nodes accounts for the word superiority effects on phoneme detection for the auditory modality. If MERGE included a visual input in its architecture, lexical nodes should receive more bottom-up support in AV than in A.

Determining the top-down or bottom-up nature of the lexical influence on phonemic process is beyond the scope of our study and further research is needed to determine how visual processing interacts with lexical activation. The timing of audiovisual integration in lexical access still remains an open question. One possibility is that the visual information directly activates lexical representations. Alternatively, the visual information could influence a pre-lexical stage. Although the present study does not provide an answer to this question, other studies with different paradigms are in progress to provide insights as to the locus of the effect of the visual information during lexical access.

Acknowledgements

We would like to thank Jean-Luc Schwartz for recording the video files and Coriandre Vilain for his very useful advice. We are also grateful to Emilie Sylvestre for proofreading this paper.

References

Amano, J., Sekiyama, K., 1998. The McGurk effect is influenced by the stimulus set size. In: Proceedings of the Auditory-Visual Speech Processing Conference. Terrigal, Australia, December 4–7, pp. 43–48.

Barutchu, A., Crewther, S., Kiely, P., Murphy, M., 2008. When /b/ill with /g/ill becomes /d/ill: evidence for a lexical effect in audiovisual speech perception. Eur. J. Cognitive Psychol. 20 (1), 1–11.

Benoît, C., Mohamadi, T., Kandel, S., 1994. Effects of phonetic context on audio-visual intelligibility of French. J. Speech Hear. Res. 37 (5), 1195–1203.

Brancazio, L., 2004. Lexical influences in audiovisual speech perception. J. Exp. Psychol. Hum. Percept. Perform. 30 (3), 445–463.

Buchwald, A.B., Winters, S.J., Pisoni, D.B., 2009. Visual speech primes open-set recognition of spoken words. Lang. Cognitive Proc. 24 (4), 580–610.

Colin, C., Radeau, M., 2003. Les illusions McGurk dans la parole: 25 ans de recherches. Ann. Psychol. 104, 497–542.

Connine, C.M., Titone, C., 1996. Phoneme monitoring. Lang. Cognitive Proc. 11 (6), 635–645.

Cutler, A., Mehler, J., Norris, D., Segui, J., 1987. Phoneme identification and the lexicon. Cognitive Psychol. 19, 141–177.

Erber, N.P., 1969. Interaction of audition and vision in the recognition of oral speech stimuli. J. Speech Hear. Res. 12 (2), 423–425.

Frauenfelder, U.H., Segui, J., Dijkstra, T., 1990. Lexical effects in phonemic processing: facilitatory or inhibitory. J. Exp. Psychol. Hum. Percept. Perform. 16 (1), 77–91.

Ganong III, W.F., 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Hum. Percept. Perform. 6 (1), 110–125.

Gow, D.W., 2003. Feature parsing: feature cue mapping in spoken word recognition. Percept. Psychophys. 65 (4), 575–590.

Green, K.P., 1998. The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In:

Appendix.

Target-present trials used in Experiment 1 (noisy conditions) and Experiment 2 (without noise). Letters in bold represent target phonemes.

Words	Phonetic form	Frequency	Pseudo-words	Phonetic form	Words	Phonetic form	Frequency	Pseudo-words	Phonetic form
a**ff**ût	[a**f**y]	1.42	a**f**é	[a**f**e]	fu**s**il	[fyzi]	48.61	fu**s**ou	[fy**z**u]
a**t**out	[a**t**u]	5.74	a**t**o	[a**t**o]	gla**c**is	[gla**s**i]	0.02	gla**ss**u	[gla**s**y]
a**v**oue	[a**v**u]	61.56	a**v**o	[a**v**o]	in**d**ou	[ẽ**d**u]	0.03	in**d**é	[ẽ**d**e]
bé**v**ue	[be**v**y]	0.36	bé**v**o	[be**v**o]	lan**d**au	[lã**d**o]	1.01	lan**d**i	[lã**d**i]
bi**s**ou	[bi**z**u]	18.4	bi**s**é	[bi**z**e]	man**t**eau	[mã**t**o]	39.97	man**t**u	[mã**t**y]
bou**ff**i	[bu**f**i]	1.16	bou**f**u	[bu**f**y]	me**ss**ie	[me**s**i]	0.69	me**ss**é	[me**s**e]
ca**d**eau	[ka**d**o]	125.79	ca**d**u	[ka**d**y]	na**z**i	[na**z**i]	7.11	na**z**é	[na**z**e]
cha**p**eau	[ʃa**p**o]	54.91	cha**p**u	[ʃa**p**y]	ni**v**eau	[ni**v**o]	50.7	ni**v**i	[ni**v**i]
choi**s**it	[ʃwa**z**i]	170.48	choi**s**é	[ʃwa**z**e]	oi**s**eau	[wa**z**o]	77.73	oi**s**i	[wa**z**i]
cla**p**ot	[kla**p**o]	0.16	cla**p**i	[kla**p**i]	per**d**u	[pɛʀ**d**y]	36.46	per**d**o	[pɛʀ**d**o]
con**f**us	[kõ**f**y]	11.02	con**f**é	[kõ**f**e]	pro**f**it	[pʀo**f**i]	14.29	pro**f**é	[pʀo**f**e]
con**v**ie	[kõ**v**i]	2.52	con**v**é	[kõ**v**e]	ré**c**it	[ʀe**s**i]	9.89	ré**ss**é	[ʀe**s**e]
cou**t**eau	[ku**t**o]	58.15	cou**t**i	[ku**t**i]	ren**d**u	[ʀã**d**y]	508.81	ren**d**eux	[ʀã**d**œ]
dé**f**i	[de**f**i]	12.24	dé**f**u	[de**f**y]	so**s**ie	[so**z**i]	5.36	so**z**ou	[so**z**u]
de**v**is	[dø**v**i]	0.94	de**v**é	[dø**v**e]	sur**t**out	[syʀ**t**u]	1.89	sur**t**i	[syʀ**t**i]
di**ss**out	[di**s**u]	3.94	di**ss**é	[di**s**e]	sur**v**ie	[syʀ**v**i]	11.64	sur**v**ou	[syʀ**v**u]
en**v**ie	[ã**v**i]	213.96	en**v**é	[ã**v**e]	vau**d**ou	[vo**d**u]	2.93	vau**d**o	[vo**d**o]

Campbell, Dodd, B., Burnham, D. (Eds.), 2004, Hearing by Eyes II: Advances in the Psychology of Speechreading and Auditory-Visual Speech. Psychology Press, Hove, England, pp. 3–26.

Kim, J., Davis, C., Krins, P., 2004. Amodal processing of visual speech as revealed by priming. Cognition 93 (1), B39–B47.

LoCasto, P.C., Connine, C.M., Patterson, D., 2007. The role of additional processing time and lexical constraint in spoken word recognition. Lang. Speech 50, 54–75.

Marslen-Wilson, W.D., 1990. Activation, competition and frequency in lexical access. In: Altmann, G.T.M. (Ed.), Cognitive Models of Speech Processing. The MIT Press, Cambridge, MA, pp. 148–172.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cognitive Psychol. 18, 1–86.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27 (2), 338–352.

New, B., Pallier, C., Ferrand, L., Matos, R., 2001. Une base de données lexicales du français contemporain sur internet: LEXIQUE. Ann. Psychol. 101, 447–462, http://www.lexique.org.

Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: feedback is never necessary. Behav. Brain Sci. 23 (3), 299–325.

Robert-Ribes, J., Lallouache, T., Escudier, P., Schwartz, J.L., 1998. Complementary and synergy in bimodal speech: auditory, visual and audiovisual identification of French oral vowels in noise. J. Acoust. Soc. Am. 103, 3677–3689.

Sams, M., Manninen, P., Surakka, V., Helin, P., Kättö, R., 1998. McGurk effect in Finnish syllabes, isolated words, and words in sentences: effects of word meaning and sentence context. Speech Commun. 26, 75–87.

Samuel, A.G., 1981. Phonemic restoration: insights from a new methodology. J. Exp. Psychol. Gen. 110, 474–494.

Spinelli, E., Ferrand, L., 2005. Psychologie du langage: l'écrit et le parlé, du signal à la signification. Armand Colin, Paris.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Summerfield, Q., 1987. Some preliminaries to a comprehensive account to audio-visual speech perception. In: Campbell, B.D.A.R. (Ed.), Hearing by Eye: The Psychology of Lipreading. Erlbaum, Londres, pp. 3–51.

Tiippana, K., Andersen, T.S., Sams, M., 2004. Visual attention modulates audiovisual speech perception. Eur. J. Cognitive Psychol. 16, 457–472.

Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167 (917), 392–393.

H. ARTICLE 2 : FORT, M., KANDEL, S., CHIPOT, J., SAVARIAUX, C., GRANJON, L. & SPINELLI, E. (EN REVISION)

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L. & Spinelli, E. (en révision). Visual speech facilitates the early phases of word recognition: Evidence from fragment priming tasks. *Language and Cognitive Processes.*

Visual speech facilitates the early phases of word recognition

Short title: visual speech primes lexical units

Visual speech facilitates the early phases of word recognition

Short title: visual speech primes lexical units

Mathilde Fort[1], Sonia Kandel[1, 2&3], Justine Chipot[1], Christophe Savariaux[3], Lionel Granjon[3] and Elsa Spinelli[1,2&4]

[1] Laboratoire de Psychologie et NeuroCognition (CNRS UMR 5105)
Université Pierre Mendès France
BP 47 - 38040 Grenoble Cedex 9
FRANCE

[2]Institut Universitaire de France
103, bd Saint-Michel
75005 Paris
FRANCE

[3] GIPSA-lab, Dpt. Parole et Cognition (CNRS UMR 5216)
Université Stendhal
BP25 - 38040 Grenoble Cedex 9
France

[4]University of California, Berkeley
USA

mathilde.fort@upmf-grenoble.fr
sonia.kandel@upmf-grenoble.fr
justine.chipot@wanadoo.fr
christophe.savariaux@gipsa-lab.grenoble-inp.fr
lionel.granjon@gipsa-lab.grenoble-inp.fr
elsa.spinelli@upmf-grenoble.fr

Mailing address:
Mathilde Fort
Université Pierre Mendès France
Laboratoire de Psychologie et NeuroCognition
BP 48, 38040 Grenoble Cedex 9, FRANCE
Tel.: +33 4.76.82.56.30; Fax : +33 4.76.82.78.34
E-mail address: mathilde.fort@upmf-grenoble.fr

ABSTRACT

When the auditory information is deteriorated by noise, watching the face of a speaker enhances speech intelligibility. Recent findings showed that decoding the facial movements of a speaker affects word recognition processes. The main objective of this study was to show that the processing of visual speech facilitates the early phases of word recognition process. We used a priming procedure paired with a lexical decision task. The primes were syllables that either shared the initial syllable with the auditory target or did not. In Experiment 1, primes were displayed in audiovisual, audio-only or visual-only conditions. There was a priming effect in all conditions. Therefore, the mere presentation of the first two phonemes – i.e., the articulatory gestures of the initial syllable– is enough visual information to activate lexical representations and initiate the word recognition process. Experiment 2 investigated the locus of the visual-only effect by examining the facilitation as a function of word frequency. The results show that the target's word frequency modulated visual-only speech priming effects. The facilitation was significant for low-frequency words. It is likely that the locus of the facilitation is not pre-lexical. We hypothesized that visual speech mostly contributes to the word recognition process when lexical access is difficult.

Keywords: visual speech, lexical access, phonological priming, lexical frequency

Visual speech facilitates the early phases of word recognition

Short title: visual speech contribution to lexical access

Evidence suggests that the visual information provided by a speaker's face enhances speech intelligibility, especially in noisy environments (Sumby & Pollack, 1954). A study in French showed that under noisy conditions, consonant and vocalic phonemes embedded in non-words were better identified in audiovisual rather than in auditory-only presentations (Benoît, Mohamadi & Kandel, 1994). Thus, decoding articulatory gestures in the presence of auditory information facilitates phoneme identification. However, there is scarce evidence regarding the contribution of visual information to word recognition. The purpose of the present study was to show that visual information not only facilitates phoneme identification, but also accelerates the recognition of words.

To our knowledge, four studies investigated word recognition in an audiovisual context. First, in a Finnish study, Sams, Manninen, Surakka, Helin & Kättö (1998) used the McGurk effect – a perceptual illusion in which an auditory /ba/ dubbed onto a visual /ga/ is perceived as "da" (McGurk & MacDonald, 1976)–, to compare the strength of audiovisual integration across situations in which integration would mean the creation of a non-word from two real words, or vice versa. For example, pairing an auditorily-presented word (e.g. /panu/, "**p**annu", stove) with another visually-presented word (e.g. /kanu/, "**k**annu", pitcher), resulted in the perception of a pseudo-word (e.g./**t**anu/). In another condition, an auditorily-presented pseudo-word (e.g. /**p**iili/) paired with visual presentation of another pseudo-word (e.g. /kiili/), resulted in perception of a word (e.g. /tiili/, "**t**iili", brick). Although they hypothesized a stronger McGurk effect for word responses than for pseudo-word responses, their results did not support this idea. The McGurk effect was similar for words and pseudo-word. The authors concluded that lexical knowledge did not mediate audiovisual integration, at least at the stage of phonetic processing.

In another study carried out in English, Brancazio (2004) combined the McGurk and Ganong (Ganong, 1980) paradigms. In the latter, participants are asked to identify a phoneme (e.g., /t/ or /d/) that varies along a synthezised continuum (e.g., t↔d). Typically, when stimuli in the continuum form words and non-words (e.g., "**d**ask" vs. **"t**ask") there is a systematic "perceptual shift" towards the phoneme that forms a word (e.g., the proportion of /t/ response is higher than /d/). By dubbing an auditory word (e.g., "**b**eg") onto a visual non-word (e.g., "**d**eg"), or a visual word (e.g., "**d**esk") onto an auditory non-word (e.g., "**b**esk"), Brancazio (2004) compared the strength of lexical activation across auditory and visual domains. The results revealed that the lexical bias was stronger in the visual word condition than in the auditory word condition. This suggests that lexical context not only influences auditory perception, but also influences visual speech processing during audiovisual word recognition (see also Barutchu, Crewther, Kiely & Murphy 2008; Windmann, 2004).

Recent data in French indicates that visual information on the articulatory gestures of the speaker not only facilitates phoneme detection, but also contributes to the process of word recognition (Fort, Spinelli, Savariaux & Kandel, 2010). In a phoneme monitoring task involving words and non-words presented in audio-only and audiovisual contexts, with noise masking the acoustic signal, consonant phonemes (e.g., the target /**p**/) were more quickly and more accurately recognised when they were embedded in words (e.g., /ʃa**p**o/ "chapeau", hat) than non-words (/ʃa**p**y/). Notably, when the acoustic signal was strongly deteriorated (i.e., at -18 dB), this "word superiority effect" was higher in the audiovisual than the audio-only condition. That the lexical bias was stronger in the audiovisual condition suggests that visual information associated with phoneme identity contributes to lexical activation during word recognition.

Although visual information undoubtedly carries phonemic information relevant to lexical identity, speech-reading alone does not provide sufficient information to allow normal hearing

adults to identify words reliably. What role does visual information have in word recognition? Does it accelerate the word recognition processes (and if yes, how)? Is it merely involved in pre-lexical processing or does it activate lexical representations? Priming tasks are particularly well adapted to address this question because primes activate information that can be manipulated experimentally. To our knowledge, two studies carried out in English examined this issue using a priming repetition procedure (Buchwald, Winters & Pisoni, 2009; Kim, Davis & Krins, 2004). Kim et al. (2004) displayed word or non-word primes (e.g., word "back" or non-word "scay") in visual only speech that were followed by a written or an auditory target that either matched the prime or not (e.g., "sharp" for word "back", or "nunth" for non-word "scay"). Using naming and lexical decision tasks (Experiment 1 to 3), the authors found a facilitatory priming effect in the matched condition (compared to the unmatched condition) when the stimuli were words, but not when they were non-words. With the same paradigm, Buchwald et al. (2009) reported that participants identified spoken words in noise more accurately when the words were preceded by a visual speech prime of the same word compared to a control condition. Taken together, these two studies suggest that the visual information in the word prime contributes to lexical processing by activating the linguistic forms that match the visual signal. Moreover, Kim et al. (2004) found that the whole visual-only presentation of a word *accelerates* the following recognition of the same word.

We hypothesized that this acceleration of the word recognition processes could be due to an activation of the early phases of lexical access. If so, the visual information corresponding to the initial phonemes of a word should provide enough information to activate lexical representations and thus accelerates word recognition. In a French priming study, Spinelli, Segui & Radeau (2001) showed that auditory primes consisting of the first syllable of a

disyllabic word facilitated recognition of the written word (e.g., auditory /kaʀ/ → written "CARTABLE" /kaʀtabl/, schoolbag) as compared to an unrelated condition (e.g., auditory /loẽ/ → written "CARTABLE" /kaʀtabl/). Their results revealed that the auditory syllable prime /kaʀ/ was enough information to activate the word recognition process of the word "CARTABLE". In the present research we used the same paradigm but with the visual modality. We examined whether visual information (i.e., the articulatory gestures) corresponding to the two initial phonemes of a word is enough to activate its lexical representation. Experiment 1 compared the priming effect across 3 different conditions of presentation of the prime: auditory-only, audiovisual or visual-only. We expected of course a priming effect in auditory-only and audiovisual, but also in visual-only. The objective of Experiment 2 was to specify the locus of the priming effect in the visual-only condition.

Experiment 1

Method

*Participant*s

Sixty-three native French speakers (15 men and 42 women, mean age = 22 years, ranged from 17 to 38 years) participated in the experiment. They reported no auditory or visual disorders.

Stimuli and Recording

The stimuli were 90 disyllabic French target words (e.g., /by.ʀo/ "bureau", desk, the dot indicates the syllable boundary, see Appendix A). They were selected from a French spoken word database (mean lexical frequency = 24.64 occurrences per million, LEXIQUE 2, New, Pallier, Brysbaert & Ferrand, 2004). Each word was paired with two monosyllabic primes that either shared the initial syllable (matching /by/) or did not (unrelated /fo/). For the purpose of the lexical decision task, 90 disyllabic pseudo-word targets were each paired with two primes (matching, unrelated). In order to reduce the proportion of matching items to 25 %, an additional 90 unrelated filler words and 90 unrelated filler pseudo-words were included. We used a small proportion of matching items and a very short Interval Inter-Stimuli (i.e., ISI = 50 ms) between the presentation of the prime and the target in order to minimize the involvement of conscious response strategies (e.g., Hamburger & Slowiaczek, 1996). Primes and targets were recorded separately by a linguistically trained female native speaker of French.

Procedure

Participants were tested individually in a quiet room. The video stimuli were shown at 25 frames/s and the auditory component was presented at a 44,100 Hz sampling rate. We used a phonological priming procedure with a lexical decision task. The syllabic primes were displayed either audiovisually, in audio-only or visually-only in three separate blocks. Stimuli were counterbalanced across six experimental lists so that each participant went though all the conditions (Auditory-Only, Audiovisual, Visual-Only x Matching, Unrelated), but heard each target only once. Each trial began with presentation of a prime, followed 50 ms later by the auditorily presented target item. Participants were asked to decide whether or not the target was a word as accurately and quickly as possible by pressing one of two response buttons. Participants' accuracy and response times from target onset were collected.

Results

Mean reaction times (RTs) on experimental words correct responses in the 6 conditions are presented in Figure 1. Errors (0.9 %) and RTs longer than 1500 ms (5 %) were removed. For every condition, we discarded the data above/below two standard-deviations (SD) from the mean (2.3 % of the RTs). We also discarded 2 target words because their error rate was over 30 %. A 3 (Modality: auditory-only vs. audiovisual vs. visual-only) x 2 (Prime Type: matching vs. unrelated) repeated measures ANOVA was conducted by participants (F_1) and by items (F_2).

(Please insert Figure 1 around here)

The analysis revealed a main effect of Prime Type, [$F_1(1, 62) = 64.96$, $p < .001$, $\eta^2_p = .51$, $F_2(1, 87) = 71.08$, $p < .001$, $\eta^2_p = .45$]. The interaction between Modality and Prime Type was significant [$F_1(2, 124) = 4.3$, $p < .05$, $\eta^2_p = .07$, $F_2(2, 174) = 4.23$, $p < .05$, $\eta^2_p = .05$]. Planned comparisons revealed that the priming effect (matching vs. unrelated) in the visual-only condition was significant, [$F_1(1, 62) = 3.93$, $p = .05$, $\eta^2_p = .06$, $F_2(1, 87) = 4.27$, $p < .05$, $\eta^2_p = .047$], but smaller than auditory-only and audiovisual [$F_1(1, 62) = 8.57$, $p = .005$, $\eta^2_p = .12$, $F_2(1, 87) = 7$, $p < .01$, $\eta^2_p = .075$]. No difference was found between audiovisual and auditory-only conditions [$F_1(1, 62) < 1$].

Due to the low percentage of errors (< 1%), no analyses were carried out on errors.

Discussion

This study investigated whether the visual information provided by the articulatory gestures that produced the first syllable of a word contributes to lexical activation during word recognition. The results indicate that the participants recognized words faster when they were preceded by their initial syllable than by an unrelated syllable prime. This facilitatory priming effect was significant for the audio-only and audiovisual conditions but for the visual-only as well. Seeing the speaker produce /po/ accelerates the recognition of "poney".

The facilitatory priming effect in the visual-only condition indicates that decoding the speaker's oro-facial gestures to produce the initial portion of a word will provide cues on the identity of that word. These cues are exploited during the word recognition process. In the visual-only condition, the modality of presentation of the prime was different form the modality of the presentation of the auditory target word, suggesting that speech processing is amodal. This is in line with studies that presented the whole word as primes (e.g., Kim et al., 2004). The results of our study allow us to go a step further. Experiment 1 provides evidence that the mere presentation of the first two phonemes –i.e., the articulatory gestures of the initial syllable– is enough visual information to activate a lexical representation and initiate the word recognition process. So, although lip-reading is not sufficient for understanding the complete identity of the word, our results show that seeing the speaker pronouncing the first two phonemes initiates the lexical access process. This suggests that the effect of visual speech concerns the early phases of the word recognition process.

The results also show that there was less facilitation for the visual-only condition than the audiovisual and auditory-only conditions. The facilitation for the audiovisual condition was not significantly greater than for the auditory-only condition. This visual-only < audiovisual ≈ auditory-only pattern may be explained by the fact that a visible speech gesture can match

several acoustic phonemes. For example, the facial gesture in the prime /by/ is similar for /py/. For the visual-only condition, the information carried by the prime /by/ may have activated the lexical representations of the words starting by /by/ –like "bureau" – but also by /py/ (e.g., "purée" /pyre/, puree). The visual-only primes may have activated more lexical candidates than the audiovisual and auditory-only primes, increasing the lexical competition and thus decreasing the size of the facilitation effect. For the same reason, the benefit of the visual information in a clear and audible speech signal (i.e., in the audiovisual condition) may have been too small to be significantly higher than the facilitation observed for the auditory information alone (i.e., in the auditory-only condition).

It should be pointed out that the experimental design used here does not allow us to draw conclusions regarding the locus of the facilitation (i.e., lexical or pre-lexical or else). If the visual information can activate lexical units during word recognition process, the facilitatory priming effect observed in Experiment 1 should vary as a function of high-level variables. In the current models of spoken-word recognition, it is assumed that a variable such as lexical frequency affects the activation level of the different lexical units during word recognition process (e.g., COHORT, Marslen-Wilson, 1987; TRACE, McClelland & Elman, 1986, Neighborhood Activation Model, NAM, Luce & Pisoni, 1998). For example, they account for the fact that a high-frequency word is processed faster than a low-frequency unit by assuming that a high-frequency word needs less activation than a low-frequency word to be recognized. Lexical frequency should thus affect lexical or post-lexical levels rather than pre-lexical stages of word recognition process (e.g., Dahan, Magnuson & Tanenhaus, 2001). If the locus of the facilitation observed in the Experiment 1 is lexical, we should observe a modulation of the facilitatory effect as a function of word frequency (see e.g., Forster & Davis, 1984; Dufour & Peereman, 2003). The aim of Experiment 2 was to examine the impact of lexical frequency of the target words on the visual-only priming effect observed in Experiment 1.

Experiment 2

Method

Participants

Twenty native French speakers (5 men and 15 women, mean age = 25 years, ranged from 20 to 31 years) participated in the experiment. None of them participated in Experiment 1. They reported no auditory or visual disorders.

Stimuli and recording.

Sixty disyllabic French target words (e.g., /bo.ku/ "beaucoup" a lot, see Appendix B) were associated to two monosyllabic primes: one for the matching (e.g., /bo/), and another for the unrelated condition (e.g., /ʁe/). Half of the target words were high-frequency (mean lexical frequency = 124.61 occurrences per million (opm); range: 26.8 - 626, LEXIQUE 3.71, New et al., 2004) and the other half were low-frequency words (mean lexical frequency = 0.78 opm; range: 0 - 3.65). There was a significant difference in Lexical Frequency between the low-frequency and the high-frequency words, $t(58) = 27.29$, $p < .001$. However, the difference in duration between the low-frequency and high-frequency targets was not significant $t(58) < 1$. These experimental items were matched to 60 disyllabic pseudo-words. There were also 60 unrelated filler words and 60 unrelated filler pseudo-words, to reduce the proportion of matching items to 25 %. The stimuli's recording was the same as in Experiment 1.

Procedure

We used a phonological priming procedure with a lexical decision task. The syllables primes were displayed in the VO condition. The high-frequency and low-frequency target words

were both displayed randomly. The experimental design and data-gathering were the same as in Experiment 1.

Results

Mean reaction times (RTs) on experimental words for correct responses for the four conditions are shown in Figure 2. Incorrect responses (8.7 %) and RTs longer than 1500 ms (1.5 %) were removed. We also removed two target pairs from the analysis because both low-frequency members of each pair had less than 50 % of errors. For each condition, we discarded data above / below two standard-deviations (SD) from the mean (1.9 % of the RTs). Percentages of errors for the four conditions were also computed.

Reaction times

A 2 (Prime Type: matching vs. unrelated) x 2 (Target Lexical Frequency: high-frequency vs. low-frequency words) repeated measures ANOVA was conducted by participants (F_1) and by items (F_2). Analysis of RTs revealed a main effect of Prime Type [$F_1(1, 19) = 14.33$, $p < .005$, $\eta^2_p = .43$, $F_2(1, 27) = 17.591$, $p < .005$, $\eta^2_p = .39$], and Target Lexical Frequency [$F_1(1, 19) = 75.15$, $p < .001$, $\eta^2_p = .80$, $F_2(1, 27) = 39.97$, $p < .001$, $\eta^2_p = .60$]. The interaction between Prime Type and Target Lexical Frequency was significant [$F_1(1, 38) = 7.01$, $p < .05$, $\eta^2_p = .27$, $F_2(1, 27) = 12.82$, $p < .005$, $\eta^2_p = .32$]. Planned comparisons revealed that the priming effect (matching vs. unrelated) was significant for the low-frequency [$F_1(1, 19) = 13.2$, $p < .005$, $\eta^2_p = .59$, $F_2(1, 27) = 27.1$, $p < .001$, $\eta^2_p = .50$], but not for the high-frequency target words, [both $F_s < 1$].

(Please insert Figure 2 around here)

Errors

A 2 (Prime Type: matching vs. unrelated) x 2 (Target Lexical Frequency: high-frequency vs. low-frequency words) repeated measures ANOVA was conducted by participants (F_1) and by items (F_2) on the percentage of errors. The analyses showed a main effect only of Target Lexical Frequency [*M* high-frequency = 1.1 %, *M* low-frequency = 15.5 %, $F_1(1, 19) = 39.21$, $p < .001$, $F_2(1, 29) = 17.32$, $p < .001$]. There was no significant effect of Prime Type, [*M* unrelated = 8.7 %, *M* matching = 7.8 %, $F_1(1, 19) < 1$, $F_2(1, 29) = 1.80$, $p > .05$], nor interaction [both Fs < 1].

Discussion

Experiment 2 was designed to determine the locus of the facilitatory effect observed for the presentation of the visual-only prime. As in Experiment 1, visual-only primes showing the articulation of the first two phonemes of a word facilitated its processing when presented auditorily as target. The results of Experiment 2 revealed that this facilitation effect was influenced by the word's frequency. The matching visual-only primes provided facilitation for the low-frequency words but not for high-frequency words (see e.g., Forster & Davis, 1984, for similar results in visual word recognition). We presume that the lack of facilitatory effect of visual speech primes on high-frequency words could be due to the fact that their recognition is already advantaged relative to the other lexical candidates (see also Goldinger Luce & Pisoni, 1992 for a similar claim). Because lexical frequency had an impact on visual speech priming, it is likely that the locus of the facilitation observed in Experiments 1 and 2 was not pre-lexical (e.g., Dufour & Peereman, 2003). Thus, this result shows that the visual information provided by the articulatory gestures of a syllable corresponding to the two first phonemes of a word activates the lexical representation of that word. Visual-only speech facilitates the early stages of word recognition process.

General Discussion

The main objective of this study was to examine the role of visual information in the process of word recognition and determine whether this information mediated pre-lexical or lexical processing. In Experiment 1 we used audiovisual, auditory-only and visual-only primes to investigate the early phases of word recognition. The primes were syllables that could or could not match the onset of a word. The results showed that priming the first syllable of a word facilitates its auditory recognition, irrespective of the modality of the prime. The results for the visual-only condition indicate that just seeing the production of a syllable that matches the onset of a word facilitates its recognition.

Experiment 2 shed some light onto the level of processing most affected by the visual information on the articulatory gestures with this priming procedure. In this experiment, visual-only syllables primed target words of contrasting lexical frequency. The results revealed that the priming effect was significant for the low-frequency words. Globally, these findings suggest that visual information taps into lexical rather than pre-lexical levels of processing during word recognition. This suggests that visual speech may play a significant role in word recognition especially when the lexical access process is more time consuming or constitutes a cognitive load (e.g., a low-frequency word). Visual information seems to contribute to the word recognition primarily when a lexical unit requires a large amount of activation to be recognized. However, another study found that in the presence of auditory information, visual speech seemed to be more helpful for "easy" words (high-frequency words with sparse neighborhood density[1]) than for "hard" words (low-frequency words with high neighborhood density) recognition (Kaiser, Kirk, Lachs & Pisoni, 2003). As this result might suggest that visual speech mostly contributes in word recognition process when lexical access is easy, this result seems to be contradictory with our assumption. Further work is

needed to determine whether these opposite findings are due to the presence vs. absence of auditory information (Kaiser et al., 2003, vs. the present study, respectively) before drawing any definitive conclusions.

However, current models of spoken-word recognition such as TRACE (McClelland & Elman, 1986) or MERGE (Norris, McQueen & Cutler, 2000) do not consider the contribution of visual information in the process of lexical access. Our study provides evidence that such models should incorporate the orofacial gestures as a source of information in their architecture (see Brancazio, 2004; Fort et al., 2010, for a supplementary discussion about this issue).

In sum, the first studies on audiovisual speech showed the importance of visual speech for intelligibility increase (Sumby & Pollak, 1954). Visual speech plays an important role on phoneme identification (Benoît et al., 1994). More recent research showed that the enhancement of phoneme identification contributes to word recognition processes (Brancazio 2004). Kim et al., (2004) showed that the presentation for the silent articulation of a whole word activates the lexical representation of that word. The priming procedure used in the present research provides evidence that the mere visual decoding of the initial syllable of a word is enough to activate its lexical representation. Therefore, seeing the articulatory gestures of a speaker facilitates the early phases of spoken word recognition (Jesse & Massaro, 2010). Studies conducted by Cathiard, Lallouache, Mohamadi & Abry (1995) indicate that the visual system decodes the visual information on lip movements such that we can identify a phoneme /y/ before its acoustic onset. One may hypothesize that during word recognition, this visual information lead over the acoustic signal is processed to pre-activate lexical units before the auditory information becomes available. This proposition is consistent with the idea that visual speech may play a priming role (Munhall & Tohkura, 1998) in lexical access. Further research is of course necessary to investigate this issue.

Interestingly, one might address the results from Kim, Davis & Krins (2004). In their Experiment 4, the authors found that the presentation of a visual-only prime that matched the same two first phonemes but not the same coda of the target (e.g., articulatory gestures for "back" → written target "BAND") provided inhibition as compared to a control condition (e.g., "leaf" → "BAND"). In our study however, we provided evidence for a facilitatory priming effect when presented a visual-only prime that matched the same two first phonemes of the target (e.g., /by/ → "bureau", desk) as compared to a control condition (e.g., /fo/ → "bureau"). These findings suggest that in one hand the mere presentation of the facial movements for two phonemes is enough to spread activation towards lexical units. In the other hand, the visual presentation for one mismatching consonant seems to be enough to provide inhibition. First, these findings suggest that visual-speech is fine-tuned. Taken together, these results suggest that visual-only speech can increase and/or reduce the size of the cohort by activating and/or inhibiting candidates during lexical competition.

FOOTNOTES

1. Phonological neighborhood size is defined as the number of words that differ from a given target by one phoneme substitution, addition, or deletion (e.g., Luce & Pisoni, 1998).

REFERENCES

Barutchu, A., Crewther, S., Kiely, P., Murphy, M., 2008. When /b/ill with /g/ill becomes /d/ill: evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology 20 (*1), 1–11.

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research*, *37*, 1195–1203.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *30*, 445–463.

Buchwald, A.B., Winters, S.J., & Pisoni, D.B. (2009). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes*, *24*, 580–610.

Cathiard, M.-A., Lallouache, M. T., Mohamadi, T., & Abry, C. (1995). Configurational vs. temporal coherence in audio-visual speech perception. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (Vol. 3, pp. 218-221). Stockholm: ICPhS.

Dahan, D., Magnuson, J.S. & Tanenhaus, M.K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology, 42*, 317–367.

Dufour, S., & Peereman, R. (2003). Inhibitory priming effects in auditory word recognition: When the target's competitors conflict with the prime word. *Cognition, 88*, B33–B44.

Forster, K. I., and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access, *Journal of Experimental. Psychology. Learning Memory and Cognition 10*, 680–698.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*, 525–532.

Ganong III, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*, 110–125.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*, 1211-1238.

Hamburger, M. B., & Slowiaczek, L. M. (1996). Phonological priming reflects lexical competition. *Psychonomic Bulletin & Review, 3*, 520-525.

Jesse, A., & Massaro, D. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, *72*, 209–225.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 46*(2), 390-404.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*, B39–B47.

Luce, P. A., & Pisoni, D. B. (1998). Recognising spoken words: The neighborhood activation model. *Ear and Hearing Research. 19*, 1–36.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71–102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Munhall, K.G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America*, *104*, 530–539.

New, B., Pallier, C., Brysbaert, M. & Ferrand, L. (2004). Lexique 2: A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, *36* (3), 516–524.

Norris, D., McQueen, J.M. & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299–325.

Sams, M., Manninen, P., Surakka, V., Helin, P. & Kättö, R. (1998). McGurk effect in Finnish syllabes, isolated words, and words in sentences: effects of word meaning and sentence context. *Speech Communication*, *26*, 75–87.

Spinelli, E., Segui, J., & Radeau, M. (2001). Phonological priming in spoken word recognition with disyllabic targets. *Language and Cognitive Processes*, *16*, 367–392.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language, 50*, 212-230.

APPENDIX A (Experiment 1)

Material used as Experimental Target Words, Unrelated (UnR) and Matching (M) Primes in Experiment 1.

UnR Primes	R Primes	Target Words	UnR Primes	R Primes	Target Words

APPENDIX B (Experiment 2)

Material used as Experimental HF (High-Frequency) and LF (Low-Frequency) Target, Unrelated and Matching Primes in Experiment 2. Frequency (in occurrences per million) indicates Word Frequency. Neigh. Density corresponds to the number of neighbors for each target.

Unrelated Primes	□atching Primes	HF □arget w ords	Le□ical □re□uency	□eigh. □ensity	L□ □arget w ords	Le□ical □re□uency	□eigh. □ensity
so [so □□	pa □□□□□□	paquet □□ak ɛ] □	36□9□	14□	pa □n [pa ɛ̃] □	1□12□	14□
□□□□□□□□	ba □□□□□□	ba □eau [□□□□□□	106□55□	1□	baque □□ba □ɛ□□	0□46□	1□
bu □[by] □	ga □□□□□□	ga □□au [□□□□□□	42□33□	11□	gadoue [gadu □□	0□37□	11□
□ □□□□ □□□	sa □□□□] □	sa □on [□□õ□□	37□06□	17□	sabo □[sabo □□	1□79□	16□
gu □n □[gš□□	fu □[□□□□	□□mer [□□□ □] □	35□91□	10□	fu □on □□□□õ□□	0□28□	10□
chi □ʃi] □	po □□□□□□	poser [□oze □□	73□73□	15□	po □e □□ote □□	0□04□	15□
re □[e □□	bo □□□□□□	beaucoup □□□oku □□	626□	3□	bo □et [bo ɛ□□	0□	5□
re □[e □□	bo □□□□□□	beaute [□□□□□□	68□57□	22□	baudet □□odɛ] □	0□11□	13□
me [me □□	tou □□□□] □	touche □□[t ʃe] □	49□46□	22□	□oupie [tupi □	1□5□	22□
□□[li] □	nou □nu] □	nouveau □□□uvo] □	170□28□	14□	nougat □□uga] □	0□89□	13□
min □□□ ɛ̃□□	co □□□□] □	cote [kote] □	250□51□	7□	copeau [□ □opo] □	0□ ˙□	6□
fe □[□□] □	so □□□□] □	sau □□r [so □e □□	57□89□	17□	saumon □□□□□ õ□□	3□65□	17□
mu [□ □] □	□□[di □	diner □□ine] □	84□73□	8□	d □van □□□□ã□□	1□03□	8□
cou □□□□] □	de □□□□□□	debut [□□□□] □	109□88□	8□	débi □[debi □	1□ ˙□	8□
ro □ʀo □□	se □□□□] □	serie □se ʀ□□□	33□34□	19□	séna □[sena □□	1□38□	18□
chu □ʃy] □	ma [□ □] □	ma □n [ma ɛ̃] □	265□03□	12□	mago □[mago] □	2□24□	12□
vo [vo] □	ca □□□□] □	cadeau □□□ado] □	98□09□	5□	caban [kab ã□□	0□	5□
ri □ʀi □	fa [□□□□	façon [fas ɔ̃□□	212□6□	9□	□□□□□□[□□□□□□	0□03□	10□
tan □□ã] □	bu [by] □	bu □eau [b □ʀo] □	156□68□	5□	bu □in [by ʁɛ̃□□	0□57□	5□
Pe [p ɘ] □	su [sy □□	sujet □□□ʒɛ□□	107□92□	19□	sumo [symo □□	0□88□	20□
Kin [k ɜ□□	mo [□ □] □	moment [mom ã□□	403□	13□	moka □moka □□	0□54□	12□
Kin [k ɜ□□	mo [□ □] □	monnaie [□ on ɛ□□	26□82□	13□	momie [□ omi □	2□45□	13□
je □ʒə□□	do □□□□□□	donne □[done □□	233□	9□	dope □□□ope] □	0□35□	9□
Pa □□□□] □	cou [ku □□	couteau □□□□□□] □	51□08□	7□	coupon □kup ɔ̃] □	0□51□	7□
gui [gi □	fo [□□□□	□□□□□[foli] □	122□47□	10□	□□□□□□□fote] □	0□	10□
cou □□□□] □	mi [□ □□□	□ □□□[mid □□	35□ ˙5□	12□	□ □□□mika] □	0□34□	12□
□□□ʀy] □	□□[□□□	ti e □□t ʁe] □	113□ ˙1□	25□	titan □ti ã□□	1□06□	14□
ou □ʒ□] □	□□[ni □	ni □eau [nivo] □	45□46□	13□	ni □he □[n ʃe] □	0□35□	13□
□□□ʀy] □	ve □□□□□□	vécu □□eky] □	51□ ˙4□	23□	vé □n [ve ʀɛ̃□□	0□05□	25□
□□□ʀy] □	ve □□□□□□	□□□□[□elo □□	32□95□	24□	□□□□□□[ve ʁ □] □	0□66□	28□

FIGURE 1

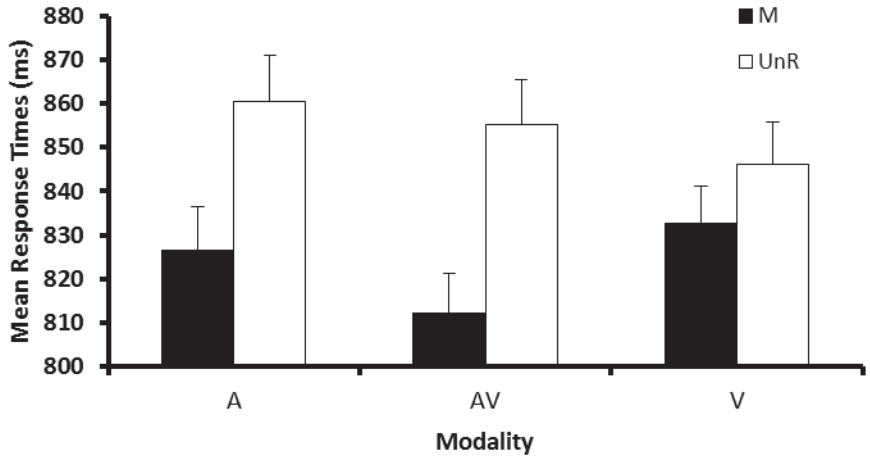


Fig. 1: Mean Response Times (in milliseconds) as a function of Prime Type (Matching vs. Unrelated) and Modality (Auditory (A) vs. Audiovisual (AV) vs. Visual-only (V)). Error bars represent Mean Standard Error.

FIGURE 2

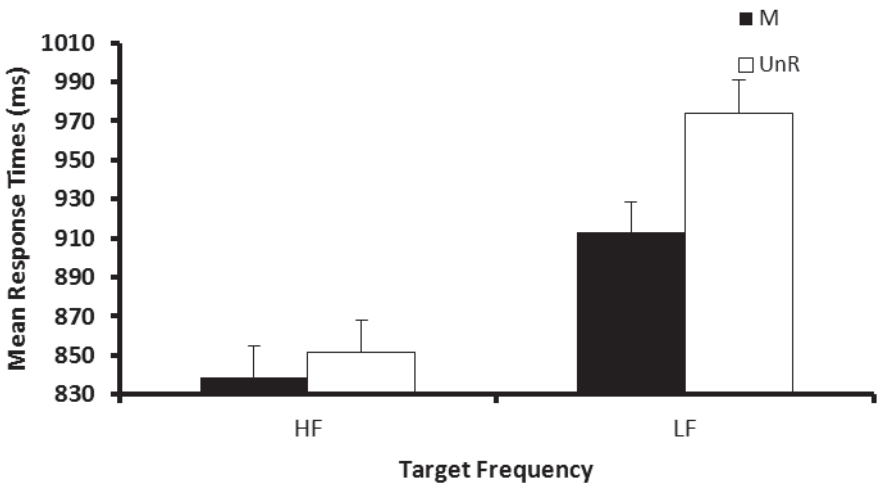


Fig. 2: Mean Response Times (in milliseconds) as a function of Prime Type (Matching (M) vs. Unrelated (UnR)) and Target Lexical Frequency (low-frequency (LF) vs. high-frequency (HF)) for the visual-only primes. Error bars represent Mean Standard Error.

I. ARTICLE 3 : FORT, M., SPINELLI, E., SAVARIAUX, C. & KANDEL, S. (RÉVISION).

Fort, M., Spinelli, E., Savariaux, C. & Kandel, S. (en révision). Audiovisual word recognition in children. *International Journal of Behavioral Development.*

Audiovisual word recognition in children

Audiovisual word recognition in children

Abstract

The goal of this study was to explore whether viewing the speaker's articulatory gestures contributes to word recognition in children (ages 5 to 10). We conducted a phoneme monitoring task with words and pseudo-words in Audio Only (AO) and Audiovisual (AV) contexts with white noise (- 9 dB) masking the acoustic signal. The results indicated that children clearly benefited from visual speech since age 5. However, the word superiority effect was not greater in the AV than the AO condition, suggesting that visual speech mostly contributes in phonemic -rather than lexical- processing.

Keywords: audiovisual speech perception; lexical access; word superiority effect, word recognition, speech development, children, noise

Several studies have shown that adults can rely on the information carried by the speaker's oro-facial gestures to identify speech in noisy situations (Benoît, Mohamadi & Kandel, 1994; Erber, 1969; Sumby & Pollack, 1954; see also Green, 1998 for a review). For instance, it has been shown that when the auditory information was deteriorated by white noise, consonant and vocalic phonemes embedded in VCVCVC nonsense words were better identified in audiovisual than in auditory-only presentations (Benoît et al., 1994). Thus, decoding facial gestures enhances phoneme intelligibility when the auditory information is deteriorated. However, little is known about the contribution of visual information to word recognition in adults (e.g., Brancazio, 2004; Fort et al. 2010) and to our knowledge, this question has never been addressed in children. The purpose of the present research was to investigate whether children use visual information in lexical access and if so at what age it starts.

Audiovisual speech perception and lexical access in adulthood

Most studies have addressed the role of visual cues in audiovisual speech perception and word recognition processes separately. Little is known about the interaction between these two sources of information. If visual speech benefits phoneme perception, it should also benefit word recognition in audiovisual face to face situations. To our knowledge, five studies have investigated word recognition in adults in an audiovisual context. A series of studies used the McGurk effect (McGurk & Mac Donald, 1976). The McGurk effect is a perceptual illusion in which an auditory /ba/ dubbed onto a visual /ga/ is perceived as /da/ or /θa/. This finding provided strong evidence that acoustic and visual signals integrate. Using this perceptual illusion, Brancazio (2004), Windmann (2004) and Barutchu, Crewther,

Kiely, Murphy & Crewther (2008) showed that visual information contributed to lexical access whereas Sams, Manninen, Surakka, Helin & Kättö (1998) were unable to find this pattern of results. More recent data from a phoneme monitoring experiment in French indicates that the presence of visual information not only facilitates phoneme detection, but also contributes to the process of word recognition in noise (Fort, Spinelli, Savariaux & Kandel, 2010). Other studies using repetition priming paradigms showed that visual-only speech primes activate lexical representations (Buchwald, Winters & Pisoni, 2009; Kim, Davis & Krins, 2004).

Most of these studies thus suggest that the visual system codes information on facial movements during speech perception and that this information is exploited during word recognition processes. The purpose of the present research was to investigate this issue from a developmental perspective.

Visual speech influence in childhood

Numerous studies have shown that very young infants are sensitive to visual speech even at the first stages of development (Burnham & Dodd, 2004; Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003; Rosenblum, Schmuckler, & Johnson, 1997). Infants seem to be able to detect auditory-visual correspondences for vowels at 2 month-olds (Patterson & Werker, 2003) and they even possess the ability to integrate these two sources of information as early as 4.5 months-olds (Burnham & Dodd, 2004). Weikum et al. (2007) also reported that 4 month-olds are able to extract relevant information from a visual-only speech stream to discriminate two languages (French vs. English). Thus, visual correlates of speech seem to influence its perception even in the early phases of human development. However, sensitivity to this information does not seem to be clearly understood in the later stages of childhood. In support of this claim, McGurk & MacDonald (1976) observed in their

original study that children between 3-5 and 7-8 years exhibited a weaker McGurk effect than adults. Moreover, Massaro, Thompson, Barron, & Laren (1986) asked preschoolers and elementary school children (4-6 and 6-10 years old) to identify an auditory /ba/ dubbed onto a visual /da/. The results showed that children responses were less dominated by the visual input (i.e., lower percentage of /da/ responses) than adults, suggesting that sensitivity to visual speech increases with age (Dupont, Aubin, & Menard, 2005; Hockley & Polka, 1994; Massaro, 1984).

A recent study (Jerger, Damian, Spence, Tye-Murray & Abdi, 2009) suggested that the differences observed between infants and children sensitivity to visual speech might have been experimentally induced from varying procedures and task demands. Indeed, infants' perception has been investigated throughout indirect measures via online responses (i.e., looking times) whereas most research on children uses direct procedures via offline responses (e.g., syllable identification) which –according to the authors- require a more conscious access and more detailed visual speech representations. To test their assumption, they examined visual speech influence on phonological processing by children using an indirect approach, the multimodal picture-word naming task. The idea underlying the original picture naming task (Jerger, Martin & Damian, 2002) is that the simultaneous presentation of a congruent distractor sharing the same onset (e.g., "peach", / piːtʃ/) would facilitate the word naming process of a picture item (e.g., "pizza", /piːtza/) as compared to the simultaneous presentation of an unrelated distractor (e.g., eagle, /iːgʌl/). In this experiment, children of ages 4 to 14 years were asked to name a picture located on the speaker's chest. To test whether visual speech may influence this naming process, the participants could only hear (auditory-only condition, AO) or both hear and see

(audiovisual condition, AV) the speaker articulating either the congruent or the unrelated distractor. As expected, the authors found that children named faster the picture when hearing simultaneously a congruent distractor than an unrelated distractor, even if they were told not to pay attention to it. Interestingly, this effect was greater for the AV than AO conditions but only for the younger (4-year-olds) and the older (10-14-year-olds) children but not for the other age groups (5, 6-7 and 8-9-year-olds). These results therefore indicate that English children between 5 to 9 years old are less influenced by visual speech than adults on both indirect and direct tasks. This is in line with the studies described above. To assess whether this lack of visual influence was due to a temporary loss of speech-reading skills, the authors also administered the participants a visual-only speech-reading task. Surprisingly, they found that the speech-reading scores increased with age. The authors argue that the lack of visual speech influence within the age of 5 to 9 could be due to a period of transition (e.g., reflecting the re-organization of phonological representational knowledge) rather than a loss of visual speech processing per se (e.g., such as speech-reading skills, see Massaro et al., 1986, for such a claim).

Recent findings investigated whether English and Japanese children (aged 6, 8 and 11 years) benefited of the presence of visual facial information to perceive speech when the auditory signal is deteriorated by noise (Sekiyama & Burnham, 2008). Using a syllable identification task, their results showed that visual speech benefits especially increased with age between 6 and 8 for English participants but remained the same for Japanese children. In the earlier stages of development (i.e., 6 years) however, the results showed that the size of visual influence was very weak and equivalent for Japanese and English children. These findings indicate that since the age of 8, children are able to extract reliable information from the speaker's orofacial gestures

to enhance the intelligibility of speech sounds. Because younger English children showed a weaker visual speech influence than their elders, these data provide evidence that speech-reading ability -to perceive speech in noise- becomes more accurate with age. Moreover, this study also showed that the size of visual speech influence across age differed between the English and Japanese participants. Together, these findings clearly indicate that language experience has differential developmental and cross-linguistic impacts on AV speech processing. Nonetheless, the mechanisms underlying the development of this capacity are still not well understood hood (e.g., Jerger, Damian, Spence, Tye-Murray & Abdi, 2009; Sekiyama & Burnham, 2008) and little is known about the specific contribution of visual processing in children during the word recognition process.

Lexical access in adults

The realization of a word can vary throughout many different factors (e.g., such as speaker, speaking rate, contexts, presence of noise, etc). Successful word recognition is thus a challenging and complex issue for novice speech perceivers. To be able to map these different realizations onto the same meaning, it is generally assumed that the mature speaker (and perceiver) has a mental lexicon (Treisman, 1960) which contains a representation of each word he/she knows (but see Goldinger, 1998; Johnson, 2007 for alternative hypotheses). Findings in adults such as the Word Superiority Effect (Cutler, Mehler, Norris & Segui, 1987), Ganong effect (Ganong, 1980) or Phonemic Restoration (e.g. Samuel, 1981), suggest that lexical representations influence phoneme perception. Cutler et al. (1987) observed that a consonant (e.g. /b/) was detected faster in a French word (e.g. “belle”, /bɛl/ i.e. beautiful) than in a pseudo-word (e.g. “berre”, /bɛʀ/). This “Word Superiority Effect”

suggests that lexical knowledge biases adult's phoneme perception/decision. However, little is known about the influence of word representations in children speech perception.

Lexical access in children

The emergence and degree of phonological specification of word representations have been investigated in toddlers (Fennell & Werker, 2003; Hallé & de Boysson-Bardies, 1996; Stager & Werker, 1997; Swingley & Aslin, 2000, 2002; see also Best, Tyler, Gooding, Orlando, & Quann, 2009, for recent findings about this issue) but only a few studies (Ackroff, 1981; Walley, 1988) examined the influence of lexical knowledge on spoken word recognition and phonological processing in older children. Both of them used the phoneme restoration paradigm (Warren, 1970). When a portion of a word corresponding to a phoneme is replaced by white noise, adult listeners tend to hear the word as intact as when white noise is added to it; they "restore" the missing speech segment. The restoration effect is greater in words (e.g., "pr**o**gress", /progʀɛs/ where the bold letter indicates the missing phoneme) than in pseudo-words (e.g., "cr**o**gless", /krogl**ɛ**s/), suggesting that a lexical bias is responsible for this effect in adults (Samuel, 1981). Since 5 (Walley, 1988), 6 and 8 year-olds (Ackroff, 1981) seem to experience less restoration than adults, we may hypothesize that lexical knowledge has less influence on phoneme perception in children than adults.

In sum, infant are able to process facial speech gestures since the first stages of their development (e.g., Kuhl & Meltzoff, 1982; Patterson & Werker, 2003) but this influence appears to be weak later on in childhood (Dupont, Aubin, & Menard, 2005; Hockley & Polka, 1994; Jerger, et al., 2009; Massaro, 1984; Massaro, et al., 1986;

McGurk & MacDonald, 1976; Sekiyama & Burnham, 2008). In addition, there is no information available on how this information could be used in children lexical processing. The goal of the present study was to get insight into this issue from a developmental perspective.

The present study

Fort et al. (2010) conducted a phoneme monitoring task in which adults had to detect French consonant targets in word and pseudo-word that were presented in noise. The results revealed that the targets were detected better and faster in words than pseudo-words. Furthermore, this word superiority effect was more important in the audiovisual than audio-only modality. These results suggest that visual speech *contributes* in lexical access per se. The aim of the present research was to investigate whether children also exploit this information in the process of word recognition. To explore this question, we decided to use a phoneme monitoring task with vowel targets. We selected vowel targets instead of consonants for several reasons. First, vowels are entities that play a primary role in speech development (e.g., Locke, 1993).Furthermore, vowels are more salient to listeners than consonants (Ladefoged, 2001) and seem to better resist in noise masking (Nooteboom & Doodeman, 1984, cited in Cutler, Sebastián-Gallés, Soler-Vilageliu & Van Ooijen, 2000). This makes the task easier for young children and therefore more adapted for a developmental study.

We conducted a vowel phoneme monitoring task involving words and pseudo-words presented in Audio only (AO) and Audiovisual (AV) contexts, with noise (i.e., -9 dB) masking the acoustic signal. Stimuli were mixed with noise in the acoustic signal to avoid ceiling effects and to enhance the role of visual speech on phonological

processing. If the influence of lexical knowledge on phonological processing increases with age, we should observe a progressive increase of the Word Superiority Effect. Similarly, if visual speech benefits increase as a function of age, there should be an AV advantage only for the older children (Sekiyama & Burnham, 2008). In other words, according to Sekiyama and Burnham's findings, we should observe only a weak AV advantage from ages 5 to 8. Finally, if visual speech contributes to lexical activation in childhood, we should replicate Fort et al.'s (2010) findings by observing a greater word superiority effect in the AV than AO presentations.

Method

Participants

Ninety-six native French speaking children participated in the experiment, ranging in age from 5 years 2 months to 10 years 10 months. The children were distributed into five groups according to age and school year: 5-6 years, kindergarten (mean age: 5 years 8 months, N = 19), 6-7 years, first grade (mean age: 6 years 11 months, N = 18), 7-8 years, second grade (mean age: 7 years 11 months, N = 20), 8-9 years, third grade (mean age: 8 years 11 months, N = 20), and 9-10 years, fourth grade (mean age: 9 years 10 months, N = 19). They all had normal or corrected-to-normal vision and reported no auditory disorders.

Stimuli

The stimulus set was composed of 40 stimuli of dissyllabic word/pseudo-word pairs selected. We selected the items that are known at age 5. Twenty pairs were target-present trials (i.e., the target phoneme was in the carrier item, see Appendix) and 20 pairs were target-absent trials. Each pseudo-word was constructed by changing the first phoneme of the first syllable in the original word (e.g., the French word "bateau" = /bato/, boat, paired with the pseudo-word /lato/). We used this procedure to insure that the pseudo-words were closed to the original words but were nevertheless non lexical items. We decided to specifically change the first phoneme in order to keep constant across the members of each pair the consonantal environment which preceded the vowel target phoneme (e.g., the target **/o/** in the word "bat**eau**" = /bat**o**/, or in the pseudo-word /lat**o**/).

Target-present trials. For the 20 pairs of target-present trials (or carrier items), the critical target phoneme was constant between each word / pseudo-word pair and was always located at the end of the second syllable (e.g., the target **/o/** in the word "bat**eau**" = /bat**o**/, or in the pseudo-word /lat**o**/). The reason for choosing this position was that the lexical effect should be stronger than if the target appeared in the initial syllable (cf. Frauenfelder Segui & Dijkstra, 1990). We used three vowel target phonemes: three rounded vowels (/o/, /y/) and one stretched vowel (/e/).

Target-absent trials. The 20 word/pseudo-word pairs of target-absent trials (e.g., /tɔʀty/, turtle vs. /tɔʀti/) were constructed using the same phonemes as in the previous paragraph. However, these pairs were preceded (at the beginning of each block) by a non-matching target phoneme (e.g., the target **/o/** in "merci" = /mɛʀsi/ or in the pseudo-word /lɛʀsi/). The non-matching target-phoneme was carefully chosen to be acoustically and visually (i.e., articulatorily) different from the second vowel phoneme.

Stimuli recording. The stimuli and target phonemes (e.g., target **/o/** for the carrier items /ʒym**o**/ and /lym**o**/) were recorded in a sound proof room by a trained female native French speaker. We only presented the face of the talker (from her chin to eyebrows) with a green background. The recording was done with a tri-CCD SONY DXC-990P camera and an AKG C1000S microphone. The recording was digitalized with the Dps Reality v 3.1.9 software to obtain mpeg video files. In the auditory-only (AO) condition we used the soundtrack extracted from the video so that the acoustic signal was identical in the AO and audiovisual (AV) conditions. We used the Matlab 7.1 software to generate the noise and add it to each utterance. We used one noise

level or Signal to Noise Ratio[1] (i.e., -9 dB). As each utterance energy was dependent on its vowel and on its consonant type (e.g., plosive, fricative) we calculated the mean strength for each stimulus and then added white noise to keep constant the Signal to Noise Ratio throughout the duration of the stimulus. The stimuli were distributed in 2 experimental lists corresponding to the 2 presentation conditions (AO vs. AV). Each list contained 10 pairs of target-present trials and 10 pairs of target-absent trials presented in random order.

Procedure

Participants were tested individually in a room apart from their classroom inside the school. They sat at 50 centimetres from an LCD screen (Neovo 17 X-17A) in a darkened sound proof room. The video stimuli were presented at 25 frames/s with a resolution of 720 x 576 pixels. The auditory component of the stimuli was provided at a 44100 Hz sampling rate by two SONY SRS-88 speakers located on both sides of the screen. The experiment was conducted with E-Prime 2.0 software (Psychological Software Tools, Pittsburgh, PA). The experiment consisted of two different sessions separated by one or two weeks. In each session, participants had to detect the target vowel in the target item (a word or a pseudo-word). They were told that the target phoneme could or could not be in the carrier utterance. A Go/No Go response task was employed whereby participants pressed the space bar of a keyboard as quickly as possible when they perceive the target phoneme in the carrier item. To limit the cognitive load, the vowel-target type was displayed by block so that the target was always displayed auditorily once before each block. Before each trial, participant's response hand was always in the "start position", namely just above the space bar. They were instructed to not do anything if they did not hear it and used the dominant hand to give their response if the target was present.

In each session, all the carrier items were either displayed in AO (with the still face of the speaker) or in AV (with the moving face of the speaker). Thus, each item was presented twice to each participant, once in AV, once in AO. Within each block, half of the items contained the target phoneme (target-present trials) and half did not (target-absent trials). For the AV condition, the participants were told to watch and listen to the stimuli carefully. This instruction intended to avoid that the participant focused on one modality more than the other (e.g., Alsius, Navarra, Campbell, Soto-Faraco, 2005). The order of the Modality of presentation was counterbalanced between participants. A training session of 6 stimuli preceded each condition and could be repeated as many times as necessary.

Results

Mean Response Times (RTs) and correct detection scores were calculated for each participant and for each target-present item pair. Due to a great variability of responses times, no analyse was carried out on RTs. A 5 (Age: 5-6 / 6-7 / 7-8 / 8-9 and 9-10 years) x 2 (Modality: AO vs. AV) x 2 (Lexical Status: word vs. pseudo-word) mixed ANOVA was conducted both by participants (*F1*) and items (*F2*). Mean Correct Detection scores for all the conditions are shown in Figure 2.

(insert the Figure about here)

The analyses revealed a main effect of age *F1* (4, 91) = 18.58, $p < .001$, $\eta^2_{p=}$ .45; *F2* (4, 19) = 43.16, $p < .001$, $\eta^2_{p=}$.65, and a linear trend indicating that scores linearly increased with age *F1* (1, 91) = 71.61, $p < .001$, $\eta^2_{p=}$.44; *F2* (4, 19) = 155, $p < .001$, $\eta^2_{p=}$.62. The main effect of the modality was also significant *F1* (1, 91) = 31.59, $p <.001$, $\eta^2_{p=}$.26; *F2* (1, 19) = 51.97, $p < .001$, $\eta^2_{p=}$.35, suggesting that participants detected better the phonemes in AV than AO. The main effect of lexical status was also significant, *F1* (1, 91) = 53.20, $p = .001$, $\eta^2_{p=}$.37; *F2* (1, 19) = 33.91, $p < .001$, $\eta^2_{p=}$.26, indicating that participants were better at detecting phonemes embedded in words with respect to pseudo-words. No interaction between these factors was observed.

To make sure that the children did not develop any response strategies, we also computed a d' for each stimulus pair, using this formula d' = z (CD) - z (FA) in which z represents the inverse of the normal cumulative distribution and CD and FA refers respectively to the mean probability of correct phoneme detection and false alarms. A 5 (Age: 5-6 / 6-7 / 7-8 / 8-9 and 9-10 year-olds) x 2 (Modality: AO vs. AV) x 2 (Lexical Status: word vs. pseudo-word) mixed ANOVA was conducted by

participants. We replicated the results obtained by analysing the correct detection scores. At -9 dB, the analyses on d' revealed a main effect of age, $F(4, 91) = 14.88$, $p < .001$, $\eta^2_p = .40$, and a strong AV advantage, $F(1, 91) = 160.22$, $p < .001$, $\eta^2_p = .64$. There was also a main effect of lexical status, $F(1, 91) = 19.63$, $p < .001$, $\eta^2_p = .18$. No interaction between these factors was observed.

Discussion

A previous study (Fort et al, 2010) showed that adults detected phonemes faster when they were embedded in words than pseudo-words. This difference was bigger when the visual information on facial movements was available (AV condition) than when it is not available (AO condition). These data clearly indicate that when the acoustic signal is deteriorated by noise, seeing the articulatory gestures of the speaker not only enhances the intelligibility of phonemes but also contributes to lexical access. The goal of this study was to examine the influence of visual speech on word recognition processes from a developmental perspective. The present experiment was conducted in children (from 5-6 to 10 years old). We conducted a phoneme detection task in words and pseudo-words, in AO and AV modality with noise in the acoustic signal.

The results first showed the influence of age on detection scores. A significant linear trend indicated that performance increased linearly with age. Data also provided evidence that lexical knowledge affected the children's performance. Indeed, our results indicate that children performed better to detect a phoneme embedded in a word than in a pseudo-word. These findings suggest that lexical knowledge biases phoneme detection processes in children at least since the age of 5. In other words, it seems that children can rely on lexical context to enhance phoneme intelligibility.

The original finding of this study is that children have greater scores in AV than AO modalities since ages 5-6. In contrast with previous findings (e.g., Sekiyama & Burnham, 2008) we did not observe any significant increase of the AV advantage over the AO condition with age. To our knowledge, this research is the first study reporting a size-similar benefit across age of coherent visual speech on performance

in children before age 8. Indeed, these data suggest that since the age of 5-6 years old, children are able to process visual speech to compensate for the lack of information in the auditory signal. This study seems to be the first set of data showing that young children (from 5-6 to 7 -8 years old) are able to disentangle an auditory signal from a noisy background by processing the articulatory gestures of a speaker when the auditory information is degraded. The audiovisual benefit could be explained by the fact that under deteriorated acoustic conditions, visual and acoustic signals complement each other (Summerfield, 1987). The auditory information (e.g. place of articulation) that has been masked by the noise is available in the visual signal and can be recovered by seeing the lips, teeth, tongue and jaw movements (Miller and Nicely, 1955). Because this AV advantage was equivalent across ages, it is likely that children of ages 5-6 detect phonemes in adverse conditions in an adult-like fashion, at least in situations where the visual information is coherent with the auditory information presented in the speech signal.

Thus, this study indicates that children of ages 5-6 to 10 not only use acoustic cues (e.g., Allen, Wightman, Kistler & Dolan, 1989) but also visual and lexical information to disentangle speech from the masker (i.e., the background noise) to perform the task.

However, contrary to our expectations, no significant interaction was obtained between Lexical Status and Modality. Unlike adults (Fort et al., 2010), this study also indicates that even if children can both rely on lexical context and visual speech separately, they do not seem to combine these two sources of information to enhance phoneme intelligibility in noise. The non-significant interaction between modality and lexicality suggests that the lexical bias (i.e., Word Superiority Effect) is not more important when the visual information was available in the speech signal.

Consequently, it seems that visual speech only contributes to phonemic (or pre-lexical) processing until the later stages of childhood (i.e., after ages 9-10). It is likely that during this period, visual speech may only spread activation towards pre-lexical units but not lexical representations. However, further research should be done to collect online measures (i.e., response times) and to be able to directly compare performance in adults and children. Nonetheless, because our results also provided evidence that lexical knowledge biased vowel phonological processing in children, we may posit that they can process and rely separately on these two signals to perceive speech but they do not exploit them together to optimize word recognition processes. This might be due to the fact that the information that can be extracted from the visual signal is not fine-grained enough during childhood and has to become more detailed with development to spread activation to the lexical units. Indeed, if visual speech only enhances pre-lexical processing during childhood but also contributes to lexical access in adulthood, we would expect to observe this shift during adolescence. Further research is in progress to determine the time period at which visual speech starts to influence lexical processing per se.

Footnotes

1. Signal to Noise Ratio, often written S/N or SNR, is a measure of signal strength relative to background noise. The ratio is usually measured in decibels (dB). We used the following formula: SNR = 20 $\log_{10}(V_s/V_n)$ in which V_s and V_n are respectively the original signal amplitude and the noise amplitude.

References

Ackroff, J.M. (1981). The interrelationship of verbal transformations, phonemic restorations and age. In *Doctoral dissertation*, The University of Wisconsin-Milwaukee. *Dissertation Abstracts International 42*, 2106.

Allen, P., Wightman, F., Kistler, D., & Dolan, T. (1989). Frequency resolution in children. *Journal of Speech and Hearing Research*, *32*, 317-322.

Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*(9), 839-843.

Barutchu, A., Crewther, S., Kiely, P., Murphy, M., 2008. When /b/ill with /g/ill becomes /d/ill: evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology 20* (1), 1–11.

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research*, *37*, 1195-1203.

Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. (2009). Development of phonological constancy: *Toddlers' perception of native- and Jamaican-accented words. Psychological Science, 20*, 539-542.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *30*, 445-463.

Buchwald, A.B., Winters, S.J., & Pisoni, D.B. (2009). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes*, *24*, 580-610.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 44*, 209–220.

Cutler, A., Mehler, J., Norris, D., Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology, 19*, 141-177.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition, 28* (5), 746-755.

Dupont, S., Aubin, J., & Menard, L. (2005). A study of theMcGurk effect in 4- and 5-year-old French Canadian children. *ZAS Papers in Linguistics, 40*, 1-17.

Erber, N. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *Journal of Speech and Hearing Research, 12*, 423-425.

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech, 46*, 245-264.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*, 525–532.

Frauenfelder, U. H., Segui, J., Dijkstra, T. (1990). Lexical effects in phonemic processing: facilitatory or inhibitory. *Journal of Experimental Psychology: Human Perception and Performance* , *16*, 77-91.

Ganong III, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Goldinger, S. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review*, *105*, 251-279.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd and D. Burnham (Eds.), *Hearing by eyes II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 3-26), Hove, England: Psychology Press.

Hallé, P. A., & de Boysson-Bardies, B. (1996). The Format of Representation of Recognized Words in Infants' Early Receptive Lexicon. *Infant Behavior and Development, 19*, 463-481.

Hockley, N., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *Journal of the Acoustical Society of America, 96*, 3309.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*, B39-B47.

Kuhl, P., & Meltzoff, A. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138-1141.

Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: A new multimodal picture word task. *Journal of Experimental Child Psychology, 102*, 40-59.

Jerger, S., Martin, R., & Damian, M. (2002). Semantic and phonological influences on picture naming by children and teenagers. Journal *of Memory and Language, 47*, 229-249.

Ladefoged, P. (2001). *Vowels and consonants: An introduction to the sounds of language*. Oxford: Blackwell.

Locke, J. L. (1993). *The Child's Path to Spoken Language*. Cambridge, MA: Harvard University Press.

Massaro, D. (1984). Children's perception of visual and auditory speech. *Child Development, 55*, 1777-1788.

Massaro, D., Thompson, L., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41*, 93-113.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

Miller, G. A., Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338-352.

Nooteboom, S. G., & Doodeman, G. J. N. (1984). Speech quality and the gating paradigm. In M. P. R. van den Broecke & A. Cohen (Eds.), *Proceedings ofthe I I th International Congress ofPhonetic Sciences* (pp. 481-485), Dordrecht, Foris.

Patterson, M., & Werker, J. (1999). Matching phonetic information in lips and voice is robust in 4. 5-month-old infants. *Infant Behavior and Development, 22*, 237-247.

Patterson, M., & Werker, J. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*, 191-196.

Rosenblum, L., Schmuckler, M., & Johnson, J. (1997). *The McGurk effect in infants. Perception and Psychophysics, 59*, 347–357.

Samuel, A. G. (1981). Phonemic Restoration: Insights From a New Methodology. *Journal of Experimental Psychology: General, 110*, 474-494.

Sams, M., Manninen, P., Surakka, V., Helin, P., Kättö, R. (1998). McGurk effect in Finnish syllabes, isolated words, and words in sentences: effects of word meaning and sentence context. *Speech Communication, 26*, 75-87.

Sekiyama, K., & Burnham, D. (2004). Issues in the development of auditory–visual speech perception: Adults, infants, and children. In S. H. Kim & D. H. Youn (Eds.). *Proceedings of International Conference on Spoken Language Processing* (Vol. 2 pp. 1137–1140). Jeju Island, South Korea.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*, 381-382.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio–visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-51). London: Erlbaum.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 76*, 147-166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13*, 480-484.

Treisman, A. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*, 242-248.

Walley, A.C. (1988). Spoken word recognition by young children and adults. *Cognitive Development, 3,* 137-165.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392-393.

Weikum, W. Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Galles, N., & Werker, J.F. (2007). Visual language discrimination in infancy. *Science, 316*, 1159.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language, 50(2)*, 212-230.

Appendix

Material used as Experimental Target Words and Target Pseudo-Words. Letters in bold represent Target Phonemes.

Target Words	Translation	Target Pseudo-Words
Bateau [bat**o**]	Boat	Lateau [lat**o**]
Cadeau [kad**o**]	Gift	Madeau [mad**o**]
Casser [kas**e**]	To Break	Dasser [das**e**]
Chanter [ʃãt**e**]	To Sing	Panter [pãt**e**]
Chapeau [ʃap**o**]	Hat	Tapeau [tap**o**]
Début [deb**y**]	Beginning	Nébut [neb**y**]
Dessus [dəs**y**]	Above	Peussus [pəs**y**]
Donner [don**e**]	To Give	Lonner [don**e**]
Fermer [fɛʀm**e**]	To Close	Termer [tɛʀm**e**]
Garder [gaʀd**e**]	To Keep	Narder [naʀd**e**]
Gâteau [gat**o**]	Cake	Nateau [nat**o**]
Goûter [gut**e**]	Snack	Nouter [nut**e**]
Jeter [ʒət**e**]	To throw	Deter [ʒət**e**]
Laver [lav**e**]	To Wash	Daver [dav**e**]
Manteau [mãt**o**]	Coat	Ganteau [gãt**o**]
Monter [mõt**e**]	To take up	Lonter [lõt**e**]
Parler [paʀl**e**]	To Speak	Darler [daʀl**e**]
Perdu [pɛʀd**y**]	Lost	Serdu [sɛʀd**y**]
Tortue [toʀt**y**]	Turtle	Gortu [goʀt**y**]
Venue [vən**y**]	Coming	Leunu [lən**y**]

Figure

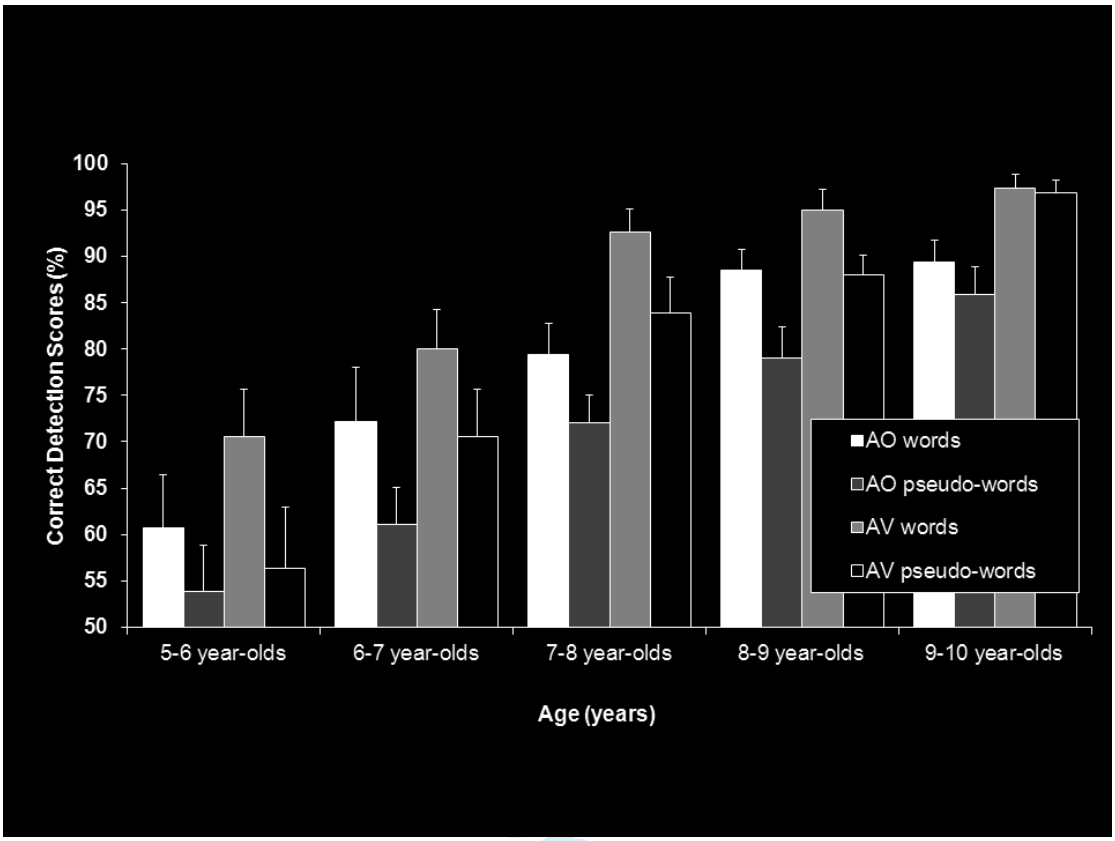
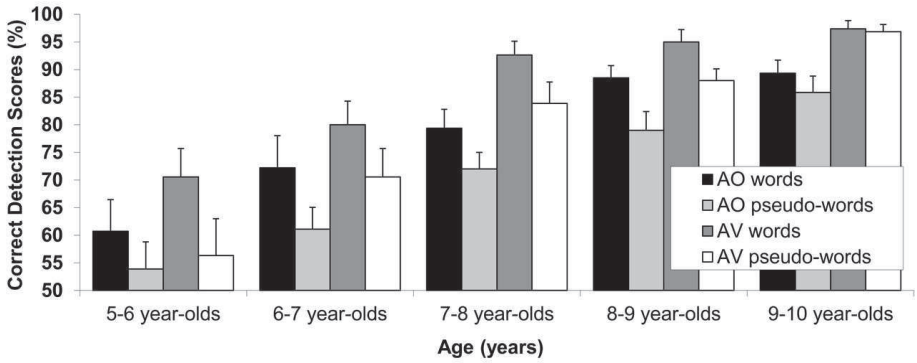


Figure. Percentages of Correct Detection Scores as a function of Age (Age: 5-6 year-olds vs. 6-7 year-olds vs. 7-8 year-olds vs. 8-9 year-olds vs. 9-10 year-olds), Modality (AO vs. AV) and Lexical Status (words vs. pseudo-words) for the -9 dB condition. Errors bars represent the standard error.

Appendix

Material used as Experimental Target Words and Target Pseudo-Words. Letters in bold represent Target Phonemes.

Target Words	Translation	Target Pseudo-Words
Bateau [bat**o**]	Boat	Lateau [lat**o**]
Cadeau [kad**o**]	Gift	Madeau [mad**o**]
Casser [kas**e**]	To Break	Dasser [das**e**]
Chanter [ʃãt**e**]	To Sing	Panter [pãt**e**]
Chapeau [ʃap**o**]	Hat	Tapeau [tap**o**]
Début [deb**y**]	Beginning	Nébut [neb**y**]
Dessus [dəs**y**]	Above	Peussus [pəs**y**]
Donner [don**e**]	To Give	Lonner [don**e**]
Fermer [fɛʀm**e**]	To Close	Termer [tɛʀm**e**]
Garder [gaʀd**e**]	To Keep	Narder [naʀd**e**]
Gâteau [gat**o**]	Cake	Nateau [nat**o**]
Goûter [gut**e**]	Snack	Nouter [nut**e**]
Jeter [ʒət**e**]	To throw	Deter [ʒət**e**]
Laver [lav**e**]	To Wash	Daver [dav**e**]
Manteau [mãt**o**]	Coat	Ganteau [gãt**o**]
Monter [mõt**e**]	To take up	Lonter [lõt**e**]
Parler [paʀl**e**]	To Speak	Darler [daʀl**e**]
Perdu [pɛʀd**y**]	Lost	Serdu [sɛʀd**y**]
Tortue [toʀt**y**]	Turtle	Gortu [goʀt**y**]
Venue [vən**y**]	Coming	Leunu [lən**y**]



Percentages of Correct Detection Scores as a function of Age (Age: 5-6 year-olds vs. 6-7 year-olds vs. 7-8 year-olds vs. 8-9 year-olds vs. 9-10 year-olds), Modality (AO vs. AV) and Lexical Status (words vs. pseudo-words) for the -9 dB condition. Errors bars represent the standard error.

115x51mm (300 x 300 DPI)

Résumé

En situation de perception audiovisuelle de la parole (i.e., lorsque deux interlocuteurs communiquent face à face) et lorsque le signal acoustique est bruité, l'intelligibilité des sons produits par un locuteur est augmentée lorsque son visage en mouvement est visible. L'objectif des travaux présentés ici est de déterminer si cette capacité à « lire sur les lèvres » nous est utile seulement pour augmenter l'intelligibilité de certains sons de parole (i.e., niveau de traitement pré-lexical) ou également pour accéder au sens des mots (i.e., niveau de traitement lexical). Chez l'adulte, nos résultats indiquent que l'information visuelle participe à l'activation des représentations lexicales en présence d'une information auditive bruitée (Etude 1 et 2). Voir le geste articulatoire correspondant à la première syllabe d'un mot constitue une information suffisante pour contacter les représentations lexicales, en l'absence de toute information auditive (Etude 3 et 4). Les résultats obtenus chez l'enfant suggèrent néanmoins que jusque l'âge de 10 ans, l'information visuelle serait uniquement décodée à un niveau pré-lexical (Etude 5).

Mots-clés : parole visuelle et audiovisuelle, reconnaissance de mots parlés, accès au lexique.

Abstract

Seeing the facial gestures of a speaker enhances phonemic identification in noise. The goal of this research was to assess whether this visual information can activate lexical representations. We investigated this question in adults (Experiment 1 to 4) and in children (Experiment 5). First, our results provide evidence indicating that visual information on consonant (Experiment 1) and vowel identity (Experiment 2) contributes to lexical activation processes during word recognition, when the auditory information is deteriorated by noise. Then, we also demonstrated that the mere presentation of the first two phonemes – i.e., the articulatory gestures of the initial syllable– is enough visual information to activate lexical representations and initiate the word recognition process (Experiment 3 and 4). However, our data suggest that visual speech mostly contributes in pre-lexical phonological -rather than lexical- processing in children till the age of 10 (Experiment 5).

Key words : speech, visual and audiovisual speech, spoken word recognition, lexical access.